Du Désordre Conformationnel des Protéines Structurées et Intrinsèquement Désordonnées par Résonance Magnétique Nucléaire

Loïc Salmon

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE
Spécialité **Physique pour les Sciences du Vivant**

Arrêté ministériel : 7 août 2006

Présentée par

Loïc SALMON

Du Désordre Conformationnel des Protéines Structurées et Intrinsèquement Désordonnées par Résonance Magnétique Nucléaire

Conformational Disorder in Folded and Intrinsically Disordered Proteins from Nuclear Magnetic Resonance

---

Thèse dirigée par **Martin BLACKLEDGE**

Thèse soutenue publiquement le **26 Novembre 2010** devant le jury composé de :

Prof. Lyndon EMSLEY	École Normale Supérieure de Lyon	**Président**
Prof. Christian GRIESINGER	Max Planck Institut Göttingen	**Rapporteur**
Prof. Bruno KIEFFER	IGBMC Strasbourg	**Rapporteur**
Prof. Hashim AL-HASHIMI	University of Michigan	**Examinateur**
Prof. Giuseppe ZACCAI	Institut Laue-Langevin	**Examinateur**
Dr. Martin BLACKLEDGE	Institut de Biologie Structurale	**Examinateur**

Thèse préparée au sein de l'Équipe Flexibilité et Dynamique des Protéines
Institut de Biologie Structurale Jean-Pierre Ebel
École Doctorale de Physique

REMERCIEMENTS

Ce manuscrit (ou tapuscrit pour utiliser un des mots les plus laids de la langue française) n'est bien sûr pas le résultat d'une recherche solitaire, mais d'un agréable et stimulant travail d'équipe. C'est pourquoi cette thèse ne serait pas complète sans ces — trop courtes — pages de remerciement.

Je souhaite en premier lieu exprimer ma reconnaissance aux membres du jury qui ont accepté de lire et juger mon travail. À Lyndon Emsley pour avoir présidé ce jury, me permettant ainsi de recevoir le titre de Docteur des mains de mon premier professeur de RMN, celui même qui me conseilla de faire cette thèse à Grenoble. Pour cela, merci encore. À Christian Griesinger, qui réussit, dans un emploi du temps plus que chargé, à trouver un créneau pour venir à ma soutenance. À Bruno Kieffer, pour avoir affronté en chemin les rigueurs du climat Grenoblois. À Guiseppe Zaccaï pour avoir subi mes diatribes philosophiques. Et enfin à Hashim Al-Hashimi, pour avoir sacrifié Thanksgiving sur l'autel de ma soutenance. Ce fut pour moi un plaisir et un honneur de soutenir ma thèse devant un tel jury et c'est pourquoi je tiens ici à tous les en remercier.

S'il est bien une personne qui doit être ici remerciée, qu'elle soit trouvée en la personne de Martin Blackledge, mon directeur de thèse et chef d'équipe. Tout d'abord pour ses indéniables qualités de scientifique, offrant toujours un regard critique, intéressé et stimulant sur l'ensemble de mes travaux mais aussi pour son très british pragmatisme venant à merveille tempérer mon — je cite — idéalisme. Enfin, pour avoir permis que cette passionnante aventure scientifique se déroule toujours dans d'excellentes conditions, en offrant un cadre de travail agréable et convivial. En un mot, merci Martin, tu resteras toujours pour moi un directeur de thèse qui n'a pas de prix.

Ces trois années de thèse n'auraient pas eu le même visage sans Malene Ringkjøbing Jensen. Tout d'abord car nous partageâmes le même bureau pendant ces trois ans. Ce fut pour moi un grand plaisir de pouvoir travailler dans ce bureau où éclats de rire et discussions scientifiques (ou non) pouvaient si bien se mélanger. Ensuite pour avoir encadrer la partie expérimentale de ma thèse, pour toujours avoir pris du temps pour m'aider ou répondre à mes questions et enfin pour avoir exploré ensemble les limites de la Thermodynamique (quand on en ajoute "un tout tout petit peu"). Tak Malene!

Je tiens de même à joindre à ces remerciements tous les membres de l'équipe avec lesquels j'ai activement travaillé. Guillaume Bouvignies, pour m'avoir initié aux secrets des GAF et à la "belle" programmation en C. Ce fut un plaisir de partager presque une année de travail, pendant laquelle tu me permis de reprendre dans les meilleures conditions les suites de ton travail. Gabrielle Nodet, certes pour tout le travail fait sur Asteroids, mais surtout pour ces improbables moments de discussion ou tout, que ce soit scientifique ou non, pouvait être passé au crible d'une analyse critique, rare de nos jours et pourtant si plaisante. Antoine Licinio, pour m'avoir produit des protéines en "Afghanistan", souvent sans aide humanitaire. Valéry Ozenne, pour assurer la relève et reprendre les rênes de l'ADMB (Association des Doctorants de Martin Blackledge). José Luis Ortéga Roldan, qui même s'il n'est pas officiellement dans l'équipe y a passé quelques mois pendant ma thèse, lançant le projet d'étude de SH3. Alexander Grimm, qui affronta sans faillir, les incessantes extrapolations nécessaires à nos scientifiques avancées. Luca Mollica, pour toujours avoir, "in case", quelques articles ou dynamiques moléculaires sous la main. Enfin Condor, certes pour son vol majestueux mais surtout pour l'efficacité avec laquelle il m'a permis de distribuer mes innombrables calculs.

Je souhaite de même remercier tous les membres du LRMN avec lesquels j'ai pu interagir de façon positive et enrichissante et particulièrement Eric Condamine et Adrien Favier pour leur assistance au spectromètre ainsi que Catherine Bougault pour avoir été ma tutrice de monitorat.

Je tiens à associer à ces remerciements les collaborateurs extérieurs au laboratoire. Concernant le développement du SF-GAF, je tiens à remercier Christian Griesinger et Rafael Brüschweiler pour leurs interrogations stimulantes concernant la quantification de la dynamique par les RDCs. Pour les travaux sur la dynamique moléculaire accélérée, je tiens à remercier Phineus Marckwick pour le vaste travail réalisé, ainsi qu'Andy McCammon. Concernant l'étude de la protéine GB3, je tiens à associer à ces remerciements Lyndon Emsley et Józef Lewandowski. Pour le projet concernant la reconnaissance moléculaire entre SH3-C et l'Ubiquitine, je tiens à remercier Nico van Nuland et José Luis Ortéga Roldan. Je tiens aussi à remercier le réseau TGIR, qui nous a permis de faire les premières mesures RMN liquide sur le spectromètre GHz du CRMN de Lyon. Enfin je souhaite remercier Edvard Munch, pour nous avoir permis d'entrevoir dans nos abysses magnétiques le spectre de son œuvre.

Il serait injuste de laisser hors de ces remerciements les enseignants, professeurs ou scientifiques qui ont jalonné mon parcours scolaire et universitaire, marquant mon évolution scientifique et me transmettant leur passion. Sans eux je n'aurais certainement pas choisi cette voie et je n'en serais

pas venu à me passionner pour des sujets si abscons pour le néophyte. Sans vouloir en dresser une liste ni exhaustive ni explicite je ne peux me priver d'en remercier quelques-uns. Du "Loup", esprit totem des merveilles redécouvertes, au "meilleur pipeteur de Bretagne", en passant par ceux qui me contèrent le mouvement des planètes, ceux qui m'initièrent à la biophysique, ou encore ceux qui osèrent m'exposer les théories de la science passée ou moderne. C'est-à-dire tous ces physiciens, chimistes, biologistes ou mathématiciens, ces épistémologues ou ces historiens des sciences qui me permirent de découvrir l'attrait de leurs disciplines, allant des bases fondamentales de la logique aux aspects les plus expérimentaux des sciences appliquées.

Par ailleurs, ce manuscrit ne serait pas ce qu'il est sans les relectures avisées qui permirent de rattraper mes égarements linguistiques. Pour cela je tiens à remercier particulièrement Martin et Malene, ainsi que Valery et Robert pour leurs contributions dans la dernière ligne droite.

Enfin je souhaite remercier tous ceux qui ont permis que ces trois années de thèse ne se résument pas à ce manuscrit.

Tout d'abord l'incroyable équipe de thésards, post-docs et assimilés qui sont passés ou sont encore au labo : Malene, Valéry, Guillaume (l'Ancien et le Jeune), Lauriane, Antoine, Robert, Luca, José Luis, Thomas (le Breton et l'Allemand), Jie-Rong, Gabrielle, Maria-Rosa, Axel, Fito, Michael, Cristina, Nico, Mathieu, Julien, Paul, Rémy, Ewen, Sophie, Joren... pour tous ces bons moments partagés hors ou au labo.

Merci à la bande de joyeux moniteurs que nous fûmes, défiant les limites de la science et de l'enseignement. Merci aux auteurs du projet X. Merci à toi Paul, pour ces discussions et pour ces randos riches en découvertes où science, philosophie et même théologie se mariaient si bien à l'air vivifiant de nos montagnes.

Je remercie aussi tous les membres de l'Aïkikaï de Grenoble : Marion, Seb, Eric, Christian, JB, Pascal, Virginie, Benoît, Ronald, Daniel, Serge et tous les autres, qui m'ont permis — modestement — de progresser dans la voie de l'Aïkido ainsi que ceux de l'antenne Jeune d'Amnesty International pour m'avoir intégré, malgré mon fonctionnement en décalé, dans cette équipe débordante d'idée et d'espoir. Merci au p'tit vélo, pour m'avoir fourni les outils nécessaires à la survie, parfois sur le fil, de mon fidèle destrier. Merci à tous ceux qui m'ont fait partager leur passion de la montagne, qui m'ont permis d'en découvrir certains aspects insoupçonnés et d'aborder au moins un instant ce monde hors du monde.

Merci à tous mes amis grenoblois. Merci à mes compagnons de l'ENS, à l'homme du Trait, à Nono et à l'auteur du Cantique des Quantiques. Merci à mes amis d'enfance que je revois avec un plaisir toujours renouvelé. Merci à toi David, pour toutes ces routes arpentées ensemble, ici ou ailleurs.

Merci à toi bergère Kazakh pour ces délicieuses lamiens, merci à toi vieil Ouighour aux toques improbables, merci à toi Juan de nous avoir guidé dans le Vilcabamba, merci à toi petit enfant si étonné de voir un étranger, merci à toi Oscar pour cette surréaliste échappée nocturne, merci à toi encore que j'oublie ou que je n'ai pas la place de remercier et qui a fait de ces trois années ce qu'elles furent.

Merci au Lapin de Tchernobyl, Nanabozo des temps modernes. Merci à toi, Olaf Tchouf, philosophe voyageur, instigateur de grandes choses en ce bas monde et fondateur du Cropisme.

Merci à toi qui un instant, un jour, ou des années, a parcouru un bout de chemin avec moi.

Mais puisque ces remerciements doivent s'achever, il ne reste plus qu'à remercier ma famille et plus particulièrement mes parents, c'est pourquoi sans vraiment échapper à la fameuse litote de basse Bretagne, je dirais juste que sans eux cela n'eut jamais été possible. Merci pour votre soutien, pour votre confiance, pour tout. Tout simplement merci.

RÉSUMÉ

Les macromolécules biologiques sont, par essence, des systèmes dynamiques. Si l'importance de cette flexibilité est maintenant clairement établie, la caractérisation précise du désordre conformationnel de ces systèmes reste encore une question ouverte. La résonance magnétique nucléaire constitue un outil unique pour sonder ces mouvements au niveau atomique que ce soit par les études de relaxation de spin ou par l'analyse des couplages dipolaires résiduels. Ces derniers permettent d'étudier l'ensemble des mouvements ayant lieu à des échelles de temps plus rapide que la milliseconde, englobant ainsi les temps caractéristiques de nombreux mouvements physiologiquement importants. L'information contenue dans ces couplages résiduels est ici interprétée principalement grâce à des approches analytiques pour quantifier la dynamique présente dans des protéines repliées, déterminer l'orientation de ces mouvements et obtenir de l'information structurale au sein de ce désordre conformationnel. Ces approches analytiques sont complémentées par des méthodes numériques, permettant ainsi soit d'observer les phénomènes sous un autre angle, soit d'examiner d'autres systèmes tels que les protéines intrinsèquement désordonnées. L'ensemble de ces études laisse transparaître une importante complémentarité entre ordre structural et désordre conformationnel.

MOTS-CLÉS Résonance Magnétique Nucléaire, Couplages Dipolaires Résiduels, Désordre Conformationnel, Protéines, Structure, Dynamique.

ABSTRACT

Biological macromolecules are, by essence, dynamical systems. While the importance of this flexibility is nowadays well established, the accurate characterization of the conformational disorder of these systems remains an important challenge. Nuclear magnetic resonance spectroscopy is a unique tool to probe these motions at atomic level, through the analysis of spin relaxation or residual dipolar couplings. The latter allows all motions occurring at timescales faster than the millisecond to be investigated, including physiologically important timescales. The information presents in those couplings is interpreted here using mainly analytical approaches in order to quantify the amounts of dynamics present in folded protein, to determine the direction of those motions and to obtain structural information within this conformational disorder. These analytical approaches are complemented by numerical methods, that allowed the observation of phenomena from a different point of view or the investigation of other systems such as intrinsically disordered proteins. All of these studies demonstrate an important complementarity between structural order and conformational disorder.

KEY WORDS Nuclear Magnetic Resonance, Residual Dipolar Coupling, Conformational Disorder, Protein, Dynamics, Structure.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

AMD	Accelerated Molecular Dynamics
ASTEROIDS	A Selection Tool for Ensemble Representations Of Intrinsically Disordered State
CS	Chemical Shift
CPMG	Carr, Purcell, Meiboom, Gill
CTAB	hexadecyl-trimethyl-ammonium bromide
DHPC	dihexanoyl-phosphatidylcholine
ditetradecyl-PC	1,2-di-O-tetradecyl-sn-glycero-3-phosphocholine
dihexyl-PC	1,2-di-O-hexyl-sn-glycero-3-phosphocholine
DNA	Deoxyribonucleic acid
DMPC	dimyristoyl-phosphatidylcholine
IDP	Intrinsically Disordered Protein
INEPT	Insensitive Nuclei Enhanced by Polarization Transfer
PAS	Principal Axis System
PEG	*n*-alkyl-poly(ethylene glycol)
PRE	Paramagnetic Relaxation Enhancement
GAF	Gaussian Axial Fluctuation
HSQC	Heteronuclear Single Quantum Correlation
LAW	Local Alignment Window
LS-GAF	Local and Shared Gaussian Axial Fluctuation
MD	Molecular Dynamics
MECCANO	Molecular Engineering Calculations using Coherent Association of Nonaveraged Orientations
MTSL	1-oxyl-2,2,5,5-tetramethyl-Δ3-pyrroline-3-methyl)-methanethiosulfonate

NMR	Nuclear Magnetic Resonance
nOe	nuclear Overhauser effect
PALES	Prediction of Alignment from Structure
RDC	Residual Dipolar Coupling
RMSD	Root Mean Square Deviation
RNA	Ribonucleic acid
S	Isotropic Motion
SCULPTOR	Structure Calculation Using Long-range, Paramagnetic, Tensorial and Orientational Restraints
SDS	sodium dodecyl-sulfate
SCRM	Self-Consistent RDC-based Model-Free approach
SECONDA	Self-Consistency of Dipolar Couplings Analysis
SF-GAF	Structure-Free Gaussian Axial Fluctuation
SVD	Singular Value Decomposition

PHYSICAL CONSTANTS

h	Planck constant	$6.626\,069\,3 \times 10^{-34}\ \mathrm{kg{\cdot}m^2{\cdot}s^{-1}}$
$\hbar$	Reduced Planck constant	$1.054\,571\,7 \times 10^{-34}\ \mathrm{kg{\cdot}m^2{\cdot}s^{-1}}$
μ_0	Permeability of vacuum	$1.256\,637\,061\,4 \times 10^{-6}\ \mathrm{kg{\cdot}m{\cdot}A^{-2}{\cdot}s^{-2}}$
g_e	Electron Landé g-factor	$-2.002\,319$
μ_B	Bohr magneton	$9.274\,009\,5 \times 10^{-24}\ \mathrm{A{\cdot}m^2}$
k_B	Boltzman constant	$1.380\,650\,5 \times 10^{-23}\ \mathrm{J{\cdot}K^{-1}}$
γ_H	^{1}H gyromagnetic ratio	$267.513 \times 10^6\ \mathrm{rad{\cdot}s^{-1}{\cdot}T^{-1}}$
γ_N	^{15}N gyromagnetic ratio	$-27.116 \times 10^6\ \mathrm{rad{\cdot}s^{-1}{\cdot}T^{-1}}$
γ_C	^{13}C gyromagnetic ratio	$67.262 \times 10^6\ \mathrm{rad{\cdot}s^{-1}{\cdot}T^{-1}}$

Part I

INTRODUCTION

INTRODUCTION

The understanding of living organisms has occupied an important place in the development of scientific and philosophic concepts due in part to its unbelievable complexity but also, and probably most importantly to the central place it occupies in the ability to humans to define themselves and to comprehend their existence.

Interestingly Biology evolved separately to Physics, which is kind of a paradox, considering that they are etymologically the Science of Life and the Science of Nature. From early Antiquity these two Disciplines were often separated into two distinct areas, Biology mainly existing as a descriptive representation of all living organism through Botany or Zoology, while Physics mainly focuses on the rationalization of inert matter based on Causality and Reproducibility of phenomena [1]. This separation, reinforced by philosophical and religious conceptions, led to a vision of the World, where a fundamental distinction was made between inert and living matter.

The real possibility to explain fundamental principles of living organism using a conceptual framework developed by physical and chemical Sciences in therefore quite recent. Even if early hypotheses had already explored this direction, such as Leonardo da Vinci drawing a parallel between combustion and nutrition, the possibility was scientifically investigated relatively recently [2]. The refutation of the so-called Vital Force Theory, that supposed the necessity of a particular force, belonging exclusively to the Living Realm, to synthesize organic compounds — organic being used in this original meaning — starts only in 1828 with the *ex vivo* synthesis of urea by Friedrich Wöher. Nevertheless this theory persisted until that the revolutionary experiments of Louis Pasteur demonstrated the nonexistence of spontaneous generation [3, 4].

Although the possibility to explore living organisms and processes using chemical and physical concepts has been revealed by numerous breakthroughs, the actual knowledge it provided was far from explaining fundamental questions about the existence, the origin and the function of living systems. The aim of Biophysics is to explore these areas. This could be seen, in my opinion, as a reductionist approach, in the non-restrictive epistemological sense, assuming that physical and chemical concepts remain valid for biological systems and can be applied to further understand Life's

phenomena. Nevertheless it should not be interpreted in terms of rigid reductionism, reducing living systems to a sum of simple physical and chemical principles. This was simply summarized by Anderson, discussing the "hierarchy" of sciences [5]:

"At each stage, entirely new laws, concepts and generalizations are necessary, requiring inspiration and creativity to just as great a degree as in the previous one. Psychology is not applied biology nor is biology applied chemistry."

The present work will join these scientific approaches by using Nuclear Magnetic Resonance (NMR) to investigate conformational flexibility of proteins.

Proteins are extremely important classes of molecules in living systems [2, 4, 6–8]. Their biological roles are extremely broad, from structural proteins, that are involved in the cell structuring, to enzymes, that catalyze biological reaction, via signaling or transport proteins, they are involved in nearly all biological processes.

Therefore the understanding of their biophysical properties is crucial for a further understanding of the function of an organism.

From a chemical point of view, a protein can be seen as a hetero-polymer comprising a combination of amino-acids fixed by the primary sequence of the considered protein. Twenty natural amino-acids, with different physico-chemical composition are encoded in the genetic code and used in protein constitution. Their order — the primary sequence — determines the intrinsic properties of the protein and thereby, its function [2, 4, 6–8].

Nevertheless proteins are biochemically very complex objects and thus their composition can not currently be used to predict and fully understand their properties [2, 4, 6–8]. During the development of molecular biology the importance of the spatial organization of the protein, its so-called fold, appeared rapidly as an important key to understand its function [9]. Three levels of organization can be distinguished. The secondary structures are structural motifs that proteins often adopt. Among them the α-helix [10] and the β-sheet [11] are the most common, but a large variety exists including different kinds of helices or turns. A second level of organization appears in the three-dimensional structure of the protein, called the tertiary structure and finally the quaternary structure corresponds to the overall organization of multimeric protein complexes.

The information derived from structural studies of proteins, as for other biological macromolecules [12], is enormous, and its main impact on molecular

interactions is formalized in the so-called "lock and key" principle developed by Emil Fisher, that assumes that interactions between two partners can be achieved if their interaction interfaces present a geometric shape complementarity. In general the paradigm of structural biology has relied very strongly on a narrow link between structure and function.

Nevertheless this static description rapidly exhibits fundamental limits that have been overcome by invoking conformational dynamics, for example in enzymatic catalysis. The dynamic behavior of proteins appears to be essential for biological function and in order to understand the underlying mechanisms, a precise characterization of the basic physics of protein motion is necessary. These dynamics can be seen as a conformational disorder, albeit occurring on a vast range of timescales, as it represents a deviation from the idealized unique and perfectly rigid description presented through static structural biology.

The accurate study of protein dynamics is experimentally challenging as it requires, in order to give a precise picture of possible motions, methods that are able to site-specifically probe biomolecular dynamics and NMR has emerged as a very well suited method for studying those motions.

One of the major strength of NMR is that it provides site-specific, even atomic resolution, information about the conformational behaviour of atoms in proteins. In fact the nucleus of an atom is characterized among other properties by its spin. Using NMR, any nucleus with a non zero spin can be studied, by inducing transition between different nuclear spin states [13–15]. Proteins are mainly made up of Hydrogen, Carbon, Nitrogen and Oxygen. The easiest species to investigate by NMR are those possessing spin-1/2 nuclei. Fortunately ^{1}H, ^{15}N, ^{13}C are spin-1/2 nuclei and therefore are well adapted for NMR studies. Due to its high gyromagnetic ratio and its natural abundance ^{1}H remains the most NMR studied nucleus but using isotopically enriched systems the low natural abundance issue of ^{15}N and ^{13}C (respectively 0.4% and 1%) can be overcome. Thus NMR studies can lead in principle to information about any H, N or C atoms present in a protein. Using appropriate NMR techniques such as high field and multi-dimensional NMR, the information of the multitude of spins can be discriminated and identified, allowing in favorable cases the characterization of each nucleus independently.

The second major interest of NMR is that NMR signals are sensitive to a very large range of timescales [15], as presented in Figure 1. Obviously all processes slower than an NMR measurement can be probed in real time, but NMR is also sensitive to motions occurring on faster timescales.

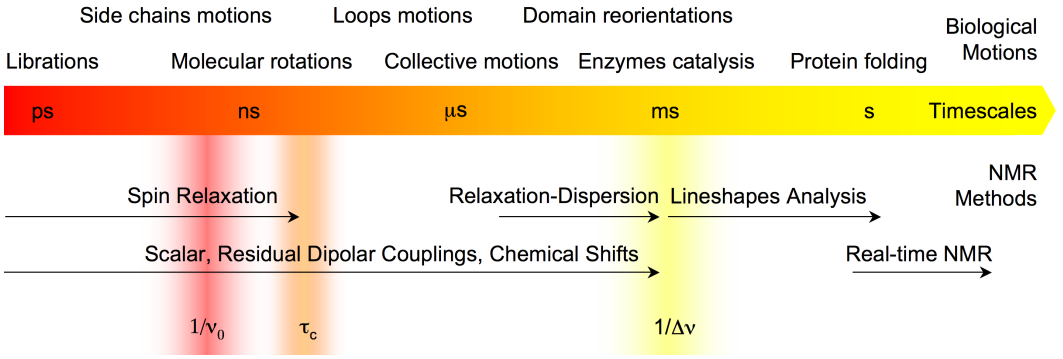

Figure 1 – NMR timescales and biological motions. Above the timescales axis: some biological motions (timescales are just indicative). Under the axis: NMR techniques to probe molecular dynamics, arrows indicate the timescales sensitivity. ν_0 represents the Larmor frequency of a studied nucleus, τ_c the correlation time of the molecule bearing this nucleus, $\Delta\nu$ is the frequency difference between two exchanging sites at which coalescence occurs. More details can be found in Chapter 1.

Very fast motions such as librations (ultra-fast bond vibrations and distortions) cannot be directly studied by NMR and they manifest themselves only as effective parameter modulations. Then all motions occurring on timescales around the Larmor frequency will appears as active in spin relaxation processes and will influence NMR spectral properties [13, 16, 17]. These dynamics can be probed by spin relaxation studies where the upper limit of accessible timescales is fixed by the overall correlation time of the molecule reorientation (5-20 ns for a medium size protein, in aqueous solution at room temperature).

Another time barrier exists in the form of chemical shift coalescence. Considering a nucleus in exchange between two chemically different sites, if the frequency of exchange is faster than this coalescence, the two resonances will merge into a single signal and for timescales not too short, this motion can be probed by relaxation-dispersion [18]. If now the exchange frequency is slower than this coalescence limit, two signals will appear and the exchange can be studied by analyzing their lineshapes.

Finally Residual Dipolar Couplings (RDCs), scalar couplings and chemical shifts are sensitive to all dynamics occurring on timescales faster than the coalescence limit [19]. They are therefore very powerful probes of biologically important motions that are thought to occur on these timescales, for example complex formation, domain reorientation, collectives motions... Even if these three interactions are sensitive to similar timescale, RDCs have been revealed to be exquisite probes of structural and dynamic information as they provide information about the orientation of internuclear vector orientations relative to the NMR static magnetic field $\mathbf{B}_0$, and therefore relative

to other bonds in the same molecule. They also present the great advantage of being measurable under different conditions, allowing the multiplication of the accessible information. For those reasons they will be at the centre of most of the analysis presented here.

The present Thesis will be organized as the following:

- A first part will focus on the theoretical concepts underlying the subsequently presented studies. A first chapter will present the main aspects of NMR Relaxation and its importance for probing biomolecular motions. The following chapter will provide a more extensive description of RDCs, which will be the main source of information in the studies presented in this work.

- The second part will focus on the development of methods to characterize protein dynamics at timescales up to the milli-second using RDCs. The first chapter will review the existing methods for studying protein dynamics using mainly RDCs and will present some approaches that will be used in the subsequent studies. The second chapter will present an analytical method for the determination of the nature and magnitude of these dynamics at atomic resolution, applied to the protein Ubiquitin. The third chapter will reinvestigate the same dynamic processes using a very complementary technique based on an accelerated molecular dynamics approach. The results obtained by the two approaches will be extensively compared. The fourth chapter will focus on the protein GB3. The analytical methods developed during the analysis of Ubiquitin will be applied to this system and the obtained dynamics, characterized by a completely local description will be revisited in order to determine the extent to which it becomes possible to reinterpret the data using a model that allows for both local and collective motions. The fifth chapter will focus on the fast and slow dynamics of an SH3 domain, for which extensive RDCs measurements have been made. The last chapter will characterize the fast dynamics of a weak complex between Ubiquitin and the same SH3 domain, using NMR relaxation.

- The last part will deal with the characterization of the unfolded states. The first chapter will briefly present unfolded proteins as highly flexible systems where some weak conformational order remains. The second chapter will focus of the extraction of local order information whereas the third will deal with the characterization of long-range order. Those studies are all based on an ensemble description of the unfolded state. Firstly, the local conformational sampling will be characterized on two systems once using RDCs and once using Chemical Shifts. Finally, the

last chapter will deal with the effect of transient long-range contacts on Paramagnetic Relaxation Enhancement and RDCs.

- Finally the Thesis will be summarized with a general concluding chapter.

Part II

THEORETICAL CONCEPTS

1

NMR RELAXATION

ABSTRACT

This chapter describes the basis of NMR relaxation that is necessary for the understanding of the work presented later in the Thesis. After a very brief introduction on NMR, the principal results of the semi-classical Redfield-Abragam theory are presented and the auto-correlation and spectral density functions are introduced. Different relaxation mechanisms that are important for ^{15}N relaxation in proteins are presented. Then emphasis is placed on the link between molecular motions and relaxation rates and finally the so-called model-free description of the spectral density function is introduced.

1.1 INTRODUCTION

NMR spectroscopy characterizes energy transitions between nuclear magnetic energy levels [13, 15–17]. As mentioned in the Introduction, one of the quantities that describes a given nucleus is the nuclear spin. This intrinsic property of the nucleus [14] is defined by a quantum number of spin s. A spin operator $\hat{\mathbf{S}}$ and an angular momentum $\hbar\hat{\mathbf{S}}$ are associated with the quantum number and can be characterized using two operators:

- $\hat{\mathbf{S}}^2$, that describes the norm of the spin angular momentum and is quantified by $s(s+1)$.
- $\hat{\mathbf{S}}_z$, which represent the component or projection of this angular momentum along an arbitrary z-axis. The corresponding quantification is obtained using m, the magnetic quantum number of spin, that can vary between $-s$ and $+s$, in steps of 1.

The magnetic moment $\hat{\boldsymbol{\mu}}_s$ associated with the angular momentum is given by:

$$\hat{\boldsymbol{\mu}}_s = \gamma\hbar\hat{\mathbf{S}} \tag{1.1}$$

where γ is the gyromagnetic ratio of the nucleus in question.

In the presence of a static magnetic field $\mathbf{B}_0$, the Hamiltonian of the interaction between the field and the magnetic moment, called the Zeeman interaction [14, 15], can be expressed as[1]:

$$\hbar\mathcal{H}^Z = -\hat{\boldsymbol{\mu}}_s.\mathbf{B}_0 = -\gamma\hbar\hat{\mathbf{S}}\mathbf{B}_0 \tag{1.2}$$

In the simple case, of an isolated spin-1/2, the Hamiltonian has two eigenstates, denoted α and β, corresponding to $m = +1/2$ and $m = -1/2$ respectively and whose energies are given by:

$$E_\alpha = -\frac{1}{2}\hbar\gamma B_0 \qquad \text{and} \qquad E_\beta = +\frac{1}{2}\hbar\gamma B_0 \tag{1.3}$$

The magnetic moment is not static but rotates around the $\mathbf{B}_0$ field with the so-called Larmor frequency [14, 15] that can be expressed in Hertz as $\nu_0 = -\gamma B_0/2\pi$ or in rad.s^{-1} as $\omega_0 = -\gamma B_0$. Transitions between the α and β states are observable at this frequency and constitute the basis of NMR spectroscopy.

More generally, as a spin system is immersed in a static magnetic field $\mathbf{B}_0$, an equilibrium occurs between the different accessible energy levels, with populations dictated by the Boltzmann distribution. Each NMR experiment consists of a perturbation of this equilibrium through the application of rf-pulses which modify the population or coherences (see below) of the spin system. The ensemble of mechanisms that will bring back the system to its initial thermal equilibrium is called relaxation. Phenomenologically two different processes are distinguished in an ensemble of isolated spins-1/2 [15]:

- the spin-lattice or longitudinal relaxation which describes the interaction of the spin system with the surrounding medium called the lattice[2] and brings back the magnetization to its equilibrium value i.e. back to the Boltzmann distribution. This process is described by a time constant, called T_1 or the associated rate $R_1 = T_1^{-1}$.

- the spin-spin or transverse relaxation that corresponds to the decay of coherences or a loss of magnetization in the plane orthogonal to $\mathbf{B}_0$.

1 In NMR the Hamiltonians are usually expressed in units of $\hbar$.

2 This lattice correspond to a thermal bath with which the system can exchange energy without modifying the temperature of the bath.

The time constant and rate associated with this decay are named T_2 and R_2 respectively.

Relaxation is of a great importance for NMR in the sense that it determines the characteristic time scale of the experiment: the delays in a given pulse sequence have to be shorter than T_2 otherwise coherences will be lost at the beginning of the acquisition, and the delay between scans has to be significantly longer than T_1 as magnetization has to return to the equilibrium value before starting a new scan.

The physical origin of relaxation can be found in the thermal motion of the considered system [13, 20]. As spin interactions are dependent on the geometry of the studied system e.g. internuclear distances, orientation towards the $\mathbf{B}_0$ field... they will be affected by any spatial modification. Thermal agitation provides a source of stochastic motion that modulates the spin interactions and therefore allows an exchange of energy between the spin system and its surroundings. The type and amount of the different motions influence the relaxation rates and relaxation can be therefore used as a probe of molecular motion.

1.2 RELAXATION THEORY

The aim of this section is to introduce the main concepts of NMR relaxation theory that will be useful for the analysis found later in this Thesis. After introducing the density matrix, the relaxation theory will be described within the framework of the semi-classical Redfield-Abragam theory before focusing on the meaning of the correlation and spectral density functions. For a more complete description of the NMR relaxation theory, the reader is referred to the references [13, 16, 17].

1.2.1 *The Density Matrix Operator*

The evolution of a spin system can be described by an operator σ called the density matrix. It is defined as the average of the operator $|\psi\rangle\langle\psi|$ over all the accessible spin states [14, 17]:

$$\sigma = \int \mathcal{P}(\psi)\,|\psi\rangle\langle\psi|\,d\tau \tag{1.4}$$

where $\mathcal{P}(\psi)$ is the density of probability of a given state $|\psi\rangle$.

Translated into a matrix description, a diagonal element of the density matrix corresponds to the population of a given eigenstate and the off-diagonal

elements correspond to the coherences between different eigenstates[3] [14, 15].

A main property of the density matrix operator is that for any experimental observable Q, the expected average value is given by:

$$\langle Q \rangle = \mathrm{Tr}\,(\sigma Q) \tag{1.5}$$

and therefore the description of the evolution of the density matrix provides a means to characterize the evolution of the whole spin system.

The time evolution of the density matrix is governed by the Liouville-von Neumann equation for a system described by a Hamiltonian $\mathcal{H}$ expressed in units of $\hbar$:

$$\frac{d\sigma(t)}{dt} = -i\left[\mathcal{H}(t), \sigma(t)\right] \tag{1.6}$$

1.2.2 *The Redfield-Abragam Theory*

The Redfield-Abragam theory is the most used in NMR and is a semi-classical description, where the spin system is treated using a quantum mechanical approach whereas the lattice is described in a classical manner. This description leads to a satisfactory characterization of relaxation processes if an *ad hoc* modification is performed in order to bring back the spin system to thermal equilibrium[4] [13, 17].

For studying relaxation processes, it is convenient to decompose the Hamiltonian in a time independent contribution $\mathcal{H}_0$ and a stochastic contribution $\mathcal{H}_1(t)$, corresponding to random fluctuation whose time average $\left\langle \mathcal{H}_1(t) \right\rangle$ is zero.

By switching to the interaction frame, which is equivalent to the classical rotating frame, and indicated here using $\tilde{}$, it is possible to remove the time dependence due to $\mathcal{H}_0$ in equation 1.6 and gives after integration:

$$\tilde{\sigma}(t) - \tilde{\sigma}(0) = -i\int_0^t \left[\tilde{\mathcal{H}_1}(t'), \tilde{\sigma}(t')\right] dt' \tag{1.7}$$

and after a second-order expansion and considering that the first-order term vanishes as $\tilde{\mathcal{H}_1}(t)$ and $\tilde{\sigma}(t)$ are assumed to be uncorrelated, and after

3 The existence of a non zero off-diagonal element means that the phases of the two involved states do not evolve completely at random but in a coherent way on average.

4 The thermal equilibrium corresponds to a situation where all coherences are lost and where populations are Boltzman weighted.

substituting $\tilde{\sigma}(0)$ with $\tilde{\sigma}(0) - \tilde{\sigma}_{eq}$ to allow return to equilibrium, we obtain:

$$\tilde{\sigma}(t) - \tilde{\sigma}(0) = -i \int_0^t \int_0^{t'} \left[\tilde{\mathcal{H}_1}(t'), \left[\tilde{\mathcal{H}_1}(t''), \tilde{\sigma}(0) - \tilde{\sigma}_{eq} \right] \right] dt'' dt' \tag{1.8}$$

As relaxation relies on the action of spin operators modulated by random spatial fluctuations it is convenient to decompose $\mathcal{H}_1(t)$ firstly, in a sum over all the different mechanisms acting on the evolution of the spin state and then in a sum of products of static spin operators $V_{\alpha,n}$ and time dependent spatial operators $F_{\alpha,n}(t)$, which describe the random fluctuations. In the following study all relaxation active operators are of second order so that they can be decomposed in the same spatial basis [21] (the second rank spherical harmonics) and therefore for $F_{\alpha,n}(t)$ the subscript n can be omitted, leading to:

$$\mathcal{H}_1(t) = \sum_n \mathcal{H}_{1,n}(t) = \sum_{\alpha,n} V_{\alpha,n} F_\alpha(t) = \sum_{\alpha,n} V^\dagger_{\alpha,n} F^*_\alpha(t) \tag{1.9}$$

where † and * denote respectively hermitian[5] and complex conjugate.

The average of the fluctuation function vanishes and those functions are assumed to be stationary, allowing the correlation function to be defined as:

$$C_{\alpha\beta}(|t - t'|) = \left\langle F_\alpha(t) F^*_\beta(t') \right\rangle \tag{1.10}$$

In the interaction representation, spin operators behave as [17]:

$$\tilde{V}_{\alpha,n}(t) = e^{\omega_\alpha t} V_{\alpha,n} \qquad \text{and} \quad \tilde{V}_{\alpha,n}(t) = e^{-\omega_\alpha t} V^\dagger_{\alpha,n} \tag{1.11}$$

where ω_α is the characteristic frequency of rotation of $V_{\alpha,n}$ in the interaction frame.

It then becomes possible to rewrite equation 1.8 as:

$$\begin{aligned} \tilde{\sigma}(t) - \tilde{\sigma}(0) = -i \sum_{\alpha,\beta,n,m} & \left[V_{\alpha,n}, \left[V^\dagger_{\beta,m}, \tilde{\sigma}(0) - \tilde{\sigma}_{eq} \right] \right] \\ & \int_0^t \int_0^{t'} C_{\alpha\beta}(|t - t'|) e^{i(\omega_\alpha t' - \omega_\beta t'')} dt'' dt' \end{aligned} \tag{1.12}$$

In this equation two different contributions can be distinguished and evaluated using the spectral density function $J_{\alpha\beta}$ defined as:

$$J_{\alpha\beta}(\omega) = \int_0^\infty C_{\alpha\beta}(\tau) e^{i\omega\tau} d\tau \tag{1.13}$$

5 Using bra-ket notation two operators A and $A^\dagger$ are hermitian conjugate one of each other if and only if for any $\langle\phi|$ and $|\psi\rangle$: $|\phi\rangle = A|\psi\rangle \Leftrightarrow \langle\phi| = \langle\psi| A^\dagger$ [14, 17].

- firstly the one corresponding to $\alpha = \beta$. Introducing $\tau = t' - t''$ and assuming that the evolution of the density matrix is much slower than the characteristic time scale of the auto-correlation function decay, the right hand side integral of equation 1.12 can be rewritten as:

$$\int_0^t \int_0^{t'} C_{\alpha\alpha}(|t' - t''|) e^{i(\omega_\alpha t' - \omega_\alpha t'')} dt'' dt' = t J_{\alpha\alpha}(\omega_\alpha) \tag{1.14}$$

- then the one with $\alpha \neq \beta$ for which the double integral of equation 1.12 can be expressed as:

$$\begin{aligned} &\int_0^t \int_0^{t'} C_{\alpha\beta}(|t' - t''|) e^{i(\omega_\alpha t' - \omega_\beta t'')} dt'' dt' \\ &\qquad = \frac{1}{i(\omega_\alpha - \omega_\beta)} \left[e^{i(\omega_\alpha - \omega_\beta)t} J_{\alpha\beta}(\omega_\beta) - J_{\beta\alpha}(\omega_\alpha) \right] \end{aligned} \tag{1.15}$$

By comparing expressions 1.14 and 1.15, for any evolution time t larger than $(\omega_\alpha - \omega_\beta)^{-1}$ all non secular terms ($\alpha \neq \beta$) can be neglected. This approximation, valid in the case of two unlike spin[6], leads for a short time t compared to the density matrix evolution to the master equation of relaxation:

$$\frac{d\tilde{\sigma}(t)}{dt} = - \sum_{\alpha,n,m} \left[V_{\alpha,n}, \left[V^\dagger_{\alpha,m}, \tilde{\sigma}(t) - \tilde{\sigma}_{eq} \right] \right] J_{\alpha\alpha}(\omega_\alpha) \tag{1.16}$$

All terms involving $n = m$ corresponds to auto-correlated processes and the others to cross-correlated phenomena, which represents interferences between two different relaxation pathways. Eventually according to equation 1.5 and using the trace property $Tr\{A[B, C]\} = Tr\{[A, B]C\}$, the master equation give for the average value of a given observable Q:

$$\begin{aligned} \frac{d\langle Q \rangle}{dt} &= Tr\left(Q \frac{d\tilde{\sigma}(t)}{dt} \right) \\ &= - \sum_{\alpha,n,m} \left\{ \left\langle \left[[Q, V_{\alpha,n}], V^\dagger_{\alpha,m} \right] \right\rangle - \left\langle \left[[Q, V_{\alpha,n}], V^\dagger_{\alpha,m} \right] \right\rangle_{eq} \right\} J_{\alpha\alpha}(\omega_\alpha) \end{aligned} \tag{1.17}$$

1.2.3 *The Auto-Correlation and Spectral Density Functions*

The auto-correlation function, introduced in the previous section, is a stationary function that characterizes the loss of memory of a randomly evolving

6 Only this case will be treated in this work, therefore this assumption will be retained.

system. It is a decreasing function whose characteristic decay time is fixed by the rate of randomly fluctuating processes.

In NMR spectroscopy, system properties are more often expressed in terms of frequency than of time evolution. This is why spectral density functions, as introduced in equation 1.13 are commonly used to describe those random processes. Following this definition the spectral density function contains a real and an imaginary part, which is often, and it will be the case in the following study, neglected. This imaginary part is responsible for the so called *dynamic frequency shifts* which brings a contribution to the chemical shift without leading to significant relaxation processes [13, 22, 23]. It is also convenient to normalize the spectral density function by the mean square spatial fluctuation leading to the following expressions [17]:

$$J(\omega) = \int_0^\infty C(\tau)\cos(\omega\tau)d\tau \qquad \text{with} \qquad C(\tau) = \frac{C_{\alpha\alpha}(\tau)}{\left\langle F_{\alpha\alpha}F^*_{\alpha\alpha}\right\rangle} \tag{1.18}$$

where J is the normalized spectral density function and C the corresponding normalized auto-correlation function. For the sake of consistency with NMR literature, these functions will simply be called the spectral density function and the correlation function.

For second-rank tensor operators the correlation function can be decomposed in the spherical harmonics basis as [24]:

$$C(t) = \frac{4\pi}{5}\sum_{m=-2}^{+2}\left\langle Y_{2m}(\vartheta(t),\varphi(t))Y_{2m}(\vartheta(0),\varphi(0))\right\rangle \tag{1.19}$$

where ϑ and φ are the polar coordinates of the vector involved in the interaction in the laboratory frame.

1.3 RELAXATION MECHANISMS

The previous section described how it is possible to derive the relaxation behavior of a given system in the presence of relaxation active interactions modulated by spatial random fluctuation. The aim of this section is to present the different kinds of interactions relevant for the description of the ^{15}N spin relaxation in biomolecular systems.

All the following interactions act as rank two tensors and can therefore be decomposed as in equation 1.9. Using this decomposition and applying the master equation for an average value (equation 1.17) to the appropriate operator (e.g. $\hat{I}_z$ for R_1 and $\hat{I}_+$ for R_2 in an isolated spin system) it becomes possible to express the relaxtion rates of NMR observables as a linear

combination of spectral density functions probed at different frequencies. Calculation details will not be developed in the following as they can be found in the literature [13, 16, 17, 24] and results will focus on the amide N_i-H_i^N spin system, which is the only experimentally studied system in this work.

1.3.1 *The Dipolar Interaction*

Each magnetic moment, such as the one induced by the spin of the nucleus, generates a magnetic field in its surroundings [13, 15, 17, 24]. Therefore, a second spin in the neighborhood of the first will experience a field dependency on their relative position. Analytical expression of this interaction can be found in Section 2.3, but it can be intuitively understood that any kind of motion that will modify the relative position of the two spins will modulate this interaction. This will represent the most important source of relaxation in amide N_i-H_i^N spin system, through the following contribution for R_1 and R_2 [24–26]:

$$R_1^D = \frac{1}{4} d_{NH}^2 \Big(3J(\omega_N) + J(\omega_H - \omega_N) + 6J(\omega_H + \omega_N) \Big) \tag{1.20}$$

$$R_2^D = \frac{1}{8} d_{NH}^2 \left(4J(0) + 3J(\omega_N) + J(\omega_H - \omega_N) + 6J(\omega_H) + 6J(\omega_H + \omega_N) \right) \tag{1.21}$$

$$\text{with} \quad d_{NH}^2 = \left(\frac{\gamma_H \gamma_N \mu_0 \hbar}{4\pi r_{NH}^3} \right)^2 \tag{1.22}$$

where r_{NH} is the internuclear distance.

Moreover the dipolar interaction is responsible for a cross-relaxation process, the nuclear Overhauser effect (nOe) [13, 15, 16]. This effect corresponds to a transfer of magnetization between two spins through space via the dipolar interaction. Therefore, if the population of a given spin is out of equilibrium, it can transfer part of its magnetization to a neighboring spin[7]. This process is often quantified by η_{NH} the ratio of intensities of an observed resonance, in presence and absence of this effect, where [24–26]:

$$\eta_{NH} = 1 + \frac{\gamma_H}{\gamma_N} \frac{d_{NH}^2 \Big(12J(\omega_H + \omega_N) - 2J(\omega_H - \omega_N) \Big)}{R_1} \tag{1.23}$$

7 This space limitation is due to the fast decrease of the interaction with the internuclear distance.

1.3.2 *The Chemical Shift Anisotropy Interaction*

Even if the isotropic chemical shifts measured in liquid state NMR do not reveal any orientational dependence, the chemical shift is an anisotropic interaction. The chemical shift translates the effect of the local electronic environment on the local magnetic field experienced by a given spin. As this distribution is *a priori* anisotropic, this will be reflected in the property of the corresponding interaction.

It has be shown experimentally that the ^{15}N amide spin chemical shift tensor σ is to a very good approximation symmetric in its principle axes frame ($\sigma_{zz} = \sigma_{\parallel}$ and $\sigma_{xx} = \sigma_{yy} = \sigma_{\perp}$) and it is commonly assumed that the z-axis of this frame is collinear with the N_i-H_i^N internuclear vector [17, 24]. This leads to the following contribution to the relaxation rates [17, 24–26]:

$$R_1^{CSA} = a_N^2 \, 6 \, J(\omega_N) \tag{1.24}$$

$$R_2^{CSA} = a_N^2 \left(4J(0) + 3J(\omega_N)\right) \tag{1.25}$$

$$\text{with} \quad a_N^2 = \frac{\left(\gamma_N B_0 (\sigma_{\parallel} - \sigma_{\perp})\right)^2}{18} \tag{1.26}$$

1.3.3 *Chemical Exchange*

Chemical exchange occurs when a given spin experiences different chemical environments [15, 16]. For example if a nucleus switches between different conformations or if a compound is involved in a chemical reaction that allows a given spin to exist in non-equivalent electronic surroundings, it will affect the NMR spectrum even if the process is at thermodynamical equilibrium as this equilibrium can be seen as a kinetic dynamic equilibrium.

Analytical descriptions for different kinds of chemical exchange can be found in the literature [16, 18, 27]. One of the simplest case [28] is, the two site exchange process where a given spin experiences only two different conformations $\mathbb{A}$ and $\mathbb{B}$, with Larmor frequencies $\omega_{\mathbb{A}}$ and $\omega_{\mathbb{B}}$, respectively. The exchange process is characterized by the following kinetic and thermodynamic parameters:

$$\mathbb{A} \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} \mathbb{B} \qquad \text{with} \qquad K = \frac{k_1}{k_{-1}} = \frac{p_{\mathbb{B}}}{p_{\mathbb{A}}} \qquad \text{and} \qquad p_{\mathbb{A}} + p_{\mathbb{B}} = 1 \tag{1.27}$$

where K is the thermodynamical constant of the equilibrium, k_1 and k_{-1} are the reaction rates and $p_{\mathbb{A}}$ and $p_{\mathbb{B}}$ are the molar fractions of the two states $\mathbb{A}$ and $\mathbb{B}$.

The exchange rate [16] is defined as $k_{ex} = k_1 + k_{-1}$ and by comparing this rate with the difference in Larmor frequencies between the two sites $\Delta\omega = |\omega_{\mathbb{A}} - \omega_{\mathbb{B}}|$, it is possible to define different regimes[8] [15, 29]:

- Slow exchange if $k_{ex} \ll \Delta\omega$. The rate of interconversion between the two states is slow enough to reveal two resonances at frequencies $\omega_{\mathbb{A}}$ and $\omega_{\mathbb{B}}$ (if one of the states is weakly populated its signal might be undetectable).

- Slow intermediate exchange when $k_{ex} < \Delta\omega$. As k_{ex} increases the number of interconversions increases leading to a broadening of both signals (the interconversion accelerates the loss of coherence in the transverse plane) and the two peaks have intermediate frequencies between $\omega_{\mathbb{A}}$ and $\omega_{\mathbb{B}}$.

- Intermediate exchange occurs at $k_{ex} \simeq \Delta\omega$. Here the two signals merge at a point called *coalescence*. In this regime the line broadening is often dramatic and can lead to a complete disappearance of any observable signal.

- Fast intermediate exchange if $k_{ex} > \Delta\omega$. In this case the number of interconversions starts to be significant: as k_{ex} increases the observed frequency gets even better averaged and the transverse relaxation gets more homogeneous resulting in resonance narrowing.

- Fast exchange when $k_{ex} \gg \Delta\omega$. Here the rate of interconversion between the two states is so fast that only an average signal can be observed. This signal is present at the frequency $\omega_{AB} = p_{\mathbb{A}}\omega_{\mathbb{A}} + p_{\mathbb{B}}\omega_{\mathbb{B}}$ and is not significantly broadened by the exchange process.

The effect of exchange in a spectrum is illustrated in Figure 2. Chemical exchange can lead to a significant increase of transverse relaxation, and this contribution is often called R_{ex}, whereas it usually results in only a small impact on the longitudinal relaxation as it is less sensitive to chemical shift changes.

1.4 MOTION ANALYSIS

One of the major interests of spin relaxation is the link between measured relaxation rates and molecular motions that include both reorientation of the entire molecule and the internal motion. Here the description of the molecular reorientation in term of tensorial behavior will be presented.

8 Often only fast, intermediate and slow exchange are defined with broaden borders.

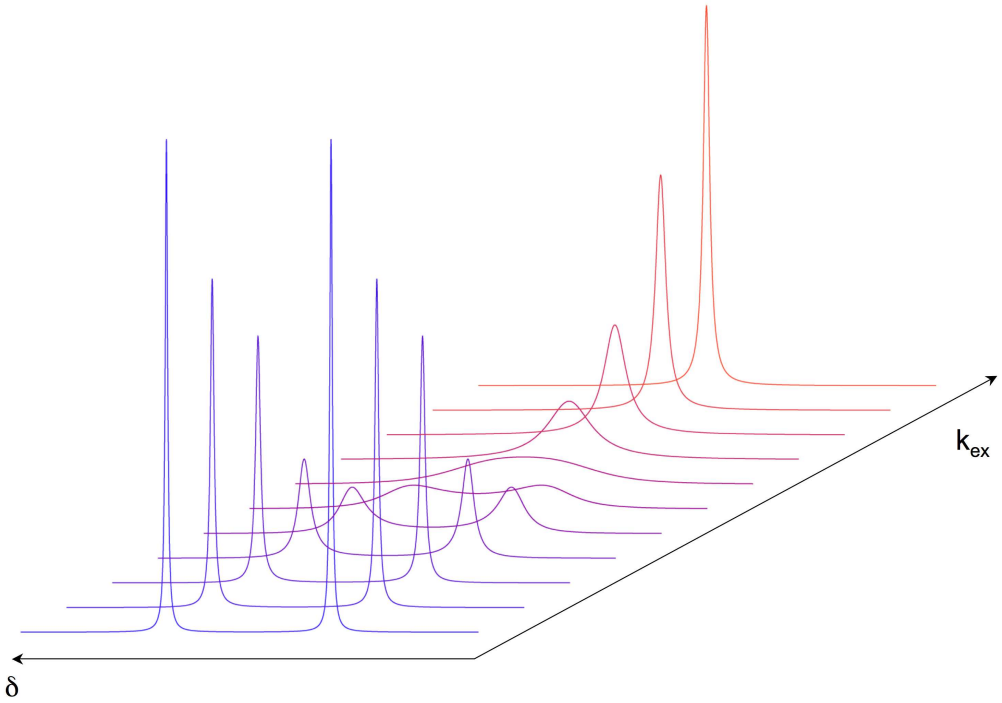

Figure 2 – Representation of chemical exchange on a spectrum. Expected spectra for two sites in exchange with equal population: exchange rates increase from blue to red.

Then two main approaches will be mentioned for describing the effect of the motions on relaxation rates, one based on a physical model of motion, the other by a generic description of the correlation function.

1.4.1 *Molecular Reorientation*

Due to the thermal agitation in liquids any molecule is subjected to constant motion. If the NMR field is homogeneous, the relaxation rates are insensitive to translational motions [15] as the spin operators active for relaxation processes have only orientational dependancies. Thus the part of the motion that has to be considered here concerns only molecular reorientations.

For a given molecule the reorientation properties are directly linked to its three-dimensional shape. If the system is not experiencing very different conformations, the description of reorientation can be done assuming a rigid structure. If the considered system presents a roughly spheric shape, the reorientation can be considered as isotropic and characterized by a single rotational diffusion coefficient. Nevertheless molecules often exhibits more complex shapes and tensorial descriptions are required. Depending on the system, an axially symmetric tensor or a fully asymmetric tensor has to be used.

The diffusion tensors can be divided into three groups: the isotropic, the axial symmetric and the fully anisotropic ones. Considering the diffusion tensor in its Principle Axis System, it can be described in terms of three eigenvalues D_{xx}, D_{yy} and D_{zz}. For the different types of tensors, the relationships between the eigenvalues are different:

- Isotropic system: $D = D_{xx} = D_{yy} = D_{zz}$.
- Axial symmetric tensor: $D_{\parallel} = D_{zz}$ and $D_{\perp} = D_{xx} = D_{yy}$. $D_{\parallel} > D_{\perp}$ corresponds to an oblate tensor, $D_{\parallel} < D_{\perp}$ to a prolate tensor.
- Fully asymmetric tensor: the three eigenvalues are different.

If the overall motion is assumed to be isotropic, the overall correlation function can be expressed as a mono-exponentially decaying process of time constant τ_c, using:

$$C_o(t) = e^{-t/\tau_c} \qquad \text{with} \qquad \tau_c = \frac{1}{6D} \tag{1.28}$$

In the case of anisotropic rotational diffusion, the correlation function can be decomposed into a multi-exponential process leading to the two following expressions [21, 24, 26, 30]:

$$C_o(t) = \sum_{r=0}^{5} A_r e^{-t/\tau_{c,r}} \tag{1.29}$$

where the coefficients A_r depend on the orientation of the considered vector (e.g. N_i-H_i^N) in the diffusion tensor principal axis frame [31]. A_r and τ_r are given by:

$$\begin{aligned}
A_1 &= 3\overline{y}^2\overline{z}^2 \qquad A_2 = 3\overline{x}^2\overline{z}^2 \qquad A_3 = 3\overline{x}^2\overline{y}^2 \\
A_{4,5} &= \frac{1}{4}\left(3(\overline{x}^4 + \overline{y}^4 + \overline{z}^4) - 1\right) \pm \frac{1}{12}\times \\
&\qquad \left(d_x(3\overline{x}^4 + 6\overline{y}^2\overline{z}^2 - 1) + d_y(3\overline{y}^4 + 6\overline{x}^2\overline{z}^2 - 1) + d_z(3\overline{z}^4 + 6\overline{y}^2\overline{x}^2 - 1)\right) \\
\tau_{c,1} &= \frac{1}{4D_{xx} + D_{yy} + D_{zz}} \qquad \tau_{c,2} = \frac{1}{D_{xx} + 4D_{yy} + D_{zz}} \\
\tau_{c,3} &= \frac{1}{D_{xx} + D_{yy} + 4D_{zz}} \qquad \tau_{c,4,5} = \frac{1}{6D_{iso} \pm 6\sqrt{D_{iso}^2 - L^2}} \\
L^2 &= \frac{D_{xx}D_{yy} + D_{xx}D_{zz} + D_{yy}D_{zz}}{3} \qquad D_{iso} = \frac{D_{xx} + D_{yy} + D_{zz}}{3} \\
d_x &= \frac{D_{xx} - D_{iso}}{\sqrt{D_{iso}^2 - L^2}} \qquad d_y = \frac{D_{yy} - D_{iso}}{\sqrt{D_{iso}^2 - L^2}} \qquad d_z = \frac{D_{zz} - D_{iso}}{\sqrt{D_{iso}^2 - L^2}}
\end{aligned} \tag{1.30}$$

where $(\overline{x}, \overline{y}, \overline{z})$ are the normalized coordinates of the considered internuclear vector.

The corresponding spectral density function, obtained by Fourier transform, is for the isotropic and anisotropic case given by:

$$J(\omega) = \frac{2}{5} \frac{\tau_C}{1 + (\omega\tau_C)^2} \tag{1.31}$$

$$J(\omega) = \frac{2}{5} \sum_{r=0}^{5} A_r \frac{\tau_{C,r}}{1 + (\omega\tau_{C,r})^2} \tag{1.32}$$

In principle reorientation tensor can be derived from hydrodynamic calculations [21, 32, 33], but it is often more convenient to derive them directly from relaxation rate analysis [34–36]. In practical, the determination of the diffusion tensor properties is made using the ratio R_2/R_1 that is not very sensitive to fast dynamics (ps-ns). This ratio can be expressed using equations 1.20, 1.21, 1.24 and 1.25 as:

$$\begin{aligned} \frac{R_2}{R_1} = & \frac{\frac{1}{8} d_{NH}^2 \left(4J(0) + 3J(\omega_N) + J(\omega_H - \omega_N) + 6J(\omega_H) + 6J(\omega_H + \omega_N)\right)}{\frac{1}{4} d_{NH}^2 \Big(3J(\omega_N) + J(\omega_H - \omega_N) + 6J(\omega_H + \omega_N)\Big) + a_N^2\, 6\, J(\omega_N)} \\ & + \frac{a_N^2 \left(4J(0) + 3J(\omega_N)\right)}{\frac{1}{4} d_{NH}^2 \Big(3J(\omega_N) + J(\omega_H - \omega_N) + 6J(\omega_H + \omega_N)\Big) + a_N^2\, 6\, J(\omega_N)} \end{aligned} \tag{1.33}$$

which makes the link between the experimentally measured rates and the spectral density function obtained by the diffusion tensor properties and the orientation of the studied vector within this tensor.

1.4.2 *Models of Local Motion*

A large number of biophysical models have been developed in order to interpret relaxation data [26, 37]. As soon as it becomes possible to derive an analytical expression of the correlation function for a vector of interest within a model, it becomes possible to establish direct connections, optimized by fitting procedures, between experimental observables and the physical properties of motion.

Proposed models include [26, 37]: rotation around an axis [38], diffusion within a cone [39], discrete jumps [40], Gaussian axial fluctuations [41, 42]... or a combination of two of them such as two-sites jumps combined

with a diffusion within a cone centered on the two sites [40, 43]. One of the main problems with those models is that it is often possible to analyze a set of experimental data with different models equally well and therefore the type of motion cannot be deduced from the experimental data, but must be chosen.

1.4.3 *Model-Free Analysis*

As a standard set of NMR relaxation experiments leads only to a small number of independent measurements, it is therefore often unrealistic to develop a complicated physical model, as this would require more parameters than those that can be fitted using the amount of information experimentally available. The analysis from Lipari and Szabo resolves this problem in analyzing relaxation data in a model-free way [18, 21, 24, 26, 44–46].

In this approach two main hypotheses are considered [44, 45]: the local and global motion are considered as statistically independent[9] and the local motion is characterized by only two physical parameters: S^2 the so-called generalized order parameter and τ_I an effective correlation time that measures the rate of decorrelation of the time dependent correlation function.

The generalized order parameter that characterizes the spatial restriction of the motion is equal to unity if the system does not have any dynamics and falls to zero if the internal motion spreads isotropically in all spatial directions. If the system presents a dynamical behavior in between those two limiting cases, the corresponding S^2 describes the asymptotic value of the correlation function and thereby how close to the initial position the system remains. This generalized order parameter can be then reinterpreted in term of different physical models.

The effective correlation time describes how fast the motion occurs and characterizes the time decay of the corresponding correlation function. The simplest model that can be assumed is a single exponential which lead for the internal correlation function to:

$$C_I(t) = S^2 + (1 - S^2)e^{-t/\tau_I} \tag{1.34}$$

9 Two motions on different timescales are sufficient to achieve the statistical independency [46].

Considering the correlation functions defined in equations 1.28 and 1.29, the resulting total correlation function is given by:

$$C(t) = C_O(t)C_I(t) \tag{1.35}$$

which corresponds (by real Fourier transform) to in the isotropic case:

$$J(\omega) = \frac{2}{5}\left[\frac{S^2\tau_C}{1+(\omega\tau_C)^2} + \frac{(1-S^2)\tau'_I}{1+(\omega\tau'_I)^2}\right] \qquad \text{with} \qquad \frac{1}{\tau'_I} = \frac{1}{\tau_C} + \frac{1}{\tau_I} \tag{1.36}$$

and in the anisotropic case to:

$$J(\omega) = \frac{2}{5}\sum_{r=0}^{5} A_r\left[\frac{S^2\tau_{C,r}}{1+(\omega\tau'_{C,r})^2} + \frac{(1-S^2)\tau'_{I,r}}{1+(\omega\tau'_{I,r})^2}\right] \qquad \text{with} \qquad \frac{1}{\tau'_{I,r}} = \frac{1}{\tau_{C,r}} + \frac{1}{\tau_I} \tag{1.37}$$

These equations have been shown to give a good agreement with experimental data if the characteristic time of the internal motion remains faster than the fastest time constant of the overall motion and if $\omega\tau_I \ll 1$.

On the one hand if τ_I is very fast and if the motion falls in the extreme narrowing limit $(\omega\tau_I)^2 \ll 1$ the previous equations can be simplified in the isotropic case as:

$$J(\omega) = \frac{2}{5}\frac{S^2\tau_C}{1+(\omega\tau_C)^2} \tag{1.38}$$

and in the anisotropic case as:

$$J(\omega) = \frac{2}{5}\sum_{r=0}^{5} A_r\left[\frac{S^2\tau_{C,r}}{1+(\omega\tau_{C,r})^2}\right] \tag{1.39}$$

On the other hand if a motion occurs without verifying the condition $\omega\tau_I \ll 1$, an extension of the model was proposed [43] by decomposing the internal motion in two different processes, one occurring on a fast time scale (corresponding to F subscripts) and one occurring at a time scale slower than the first motion but still faster than the correlation time (denoted with S subscripts), leading to the following correlation function:

$$C_I(t) = S_F^2S_S^2 + S_F^2(1-S_S^2)e^{-t/\tau_s} + (1-S_F^2)e^{-t/\tau_F} \qquad \text{and} \qquad S^2 = S_F^2S_S^2 \tag{1.40}$$

In the case of isotropic overall global motion:

$$J(\omega) = \frac{2}{5}\left[\frac{S_F^2S_S^2\tau_C}{1+(\omega\tau_C)^2} + \frac{S_F^2(1-S_S^2)\tau'_S}{1+(\omega\tau'_S)^2} + \frac{(1-S_F^2)\tau'_F}{1+(\omega\tau'_F)^2}\right] \tag{1.41}$$

$$\text{with} \qquad \frac{1}{\tau'_S} = \frac{1}{\tau_C} + \frac{1}{\tau_S} \qquad \text{and} \qquad \frac{1}{\tau'_F} = \frac{1}{\tau_C} + \frac{1}{\tau_F} \tag{1.42}$$

This model characterizes the internal motion with four parameters S_S^2, τ_S, S_F^2 and τ_F and therefore requires more information than the one provided by the standard set of experimental data. Therefore, it is usually assumed that the fast internal motion is extremely fast which removes the τ_F dependency as the previous expression simplifies to:

$$J(\omega) = \frac{2}{5}\left[\frac{S_F^2 S_S^2 \tau_C}{1+(\omega\tau_C)^2} + \frac{S_F^2(1-S_S^2)\tau'_S}{1+(\omega\tau'_S)^2}\right] \tag{1.43}$$

which for an anisotropically diffusing system gives:

$$J(\omega) = \frac{2}{5}\sum_{r=0}^{5} A_r\left[\frac{S_F^2 S_S^2 \tau_{C,r}}{1+(\omega\tau_{C,r})^2} + \frac{S_F^2(1-S_S^2)\tau_{S,r}}{1+(\omega\tau_{S,r})^2}\right] \tag{1.44}$$

1.5 CONCLUSION

NMR spin relaxation has early appeared as a very powerful probe of molecular dynamics. It was historically the main source of dynamic information for NMR spectroscopy and is nowadays still widely used. Relevance of spin relaxation to probe molecular motion come from the link between geometrical modulations — motions — and measurable relaxations rates.

Relaxation rates are sensitives to both molecule global reorientation and local motion. Both aspects are often considered as uncorrelated and it then becomes possible to express properties of reorientation in terms of a reorientation tensor. Then the local dynamics has to be described. Due to the limited number of data, the selection of a particular physical model can be difficult and thus the so-called model-free description has imposed itself as a relevant way of analyzing experimental data. This model allows to simply parameterize the spectral density function without invoking a particular model with often two parameters: one characterizing the spatial extension of the motion and one estimating the rate of decorrelation. The spectral density function then directly governs the different relaxation rates as it represents the distribution of frequencies or energies present in the lattice that are available to induce relaxation active spin transitions.

Nevertheless relaxation data analysis are sensitive only to fast dynamics. The limitation is particularly important for biological system for which most of the relevant motions are expected to be on slower timescales. Part of this

problem can be partially solved by using relaxation-dispersion experiments [18, 28, 29]. Those experiments allow to characterize exchanging system dynamics by interpreting the modulation of R_2 rates, R_{ex}, due to chemical exchange for a nucleus that experiences different chemical environments with interconversion rates on the micro to millisecond timescale. Nonetheless, a large timescale gap remains that can be conveniently probed by RDCs.

2

RESIDUAL DIPOLAR COUPLINGS

ABSTRACT

The dipolar interaction, that is partially averaged in anisotropic liquids, is an important source of structural and dynamic information for biological macromolecules. This chapter presents some experimental aspects of anisotropic liquid state NMR and derives from the expression of the dipolar interaction the analytical expressions of dynamically averaged RDCs. The structural dependency of RDCs is discussed and the averaging effect of some frequently used models of motion for RDCs analysis are derived.

2.1 INTRODUCTION

The dipolar, or dipole-dipole interaction results from the coupling of a given magnetic moment with any other magnetic moment in its surroundings [13, 15–17]. The presence of this interaction is a potentially powerful source of information for the study of the structure and dynamics of molecular systems. However as a spin, or more precisely the magnetic moment induced by this spin, can interact with all neighbouring magnetic dipoles this information can be encoded in a complex way.

Two extreme cases can be considered [13, 15]: the isotropic case where the interaction is averaged to zero and the information contents lost. This is the case in free solution and can be contrasted with the solid state case, where, in absence of macroscopic motion of the sample, the entire dipolar interaction is present. The idea of using a state of matter with an order in between the solid and the liquid state, that is to say a mesomorphic phase or liquid crystal [47], is to keep the information content of the dipolar coupling and the simplicity of liquid state NMR. The resulting partially averaged dipolar interaction will lead to the so called Residual Dipolar Couplings that retains a fraction of the full interaction strength but critically also retains the geometric dependence.

The dipolar interaction is a relatively strong interaction on the NMR scale (kHz range) [13, 15] and this strength has a direct impact on the sensitivity of the measured coupling to motions occurring on a large range of timescales. In fact any interaction is modulated by any motion occurring on timescales faster than the interaction magnitude expressed in frequency units. It can be physically understood in the following way: if a motion induce transitions between energy levels much closer in energy than the characteristic amplitude of the studied interaction (e.g. dipolar interaction), the probability and therefore the rate of transition between those energy levels will be much higher than those characterizing the larger interaction. As a consequence any measurement of energy transition between two energy levels of this large interaction will correspond to transitions between motionally averaged states.

Typical ranges of RDCs [19, 48], are dozens of Hertz which correspond to timescales of a few tens of millisecond. As the chemical shift coalescence limit often occurs on similar timescales around the millisecond[1], a RDC will be sensitive to all timescales up to the coalescence limit. This property will confer to RDCs an important role as motional probes.

Historically the NMR of solute dissolved in anisotropic media began in the 1960's with the measurement of RDCs for benzene and other small molecules in organic nematic phases [49], but the use of organic solvent and the rapid increase of spectra complexity in such highly ordered liquid crystals were suitable only for the study of small molecules. Then other small compounds were aligned in the magnetic field through their own anisotropic para- or dia-magnetic susceptibility [50]. As this kind of alignment is much weaker than the one induced by liquid crystals and does not require organic solvents, it opened the way to the first application to biologically relevant compounds [51] and then to proteins [52]. Eventually the use of dilute liquid crystals [53] where the degree of order is small enough to keep spectral resolution and large enough to give precise RDCs opened new horizons for the use of anisotropic interactions in liquid state NMR.

This chapter will first deal with experimental considerations about anisotropic liquid state NMR, before presenting the origin of RDCs, and addressing the way by which motion modulates these phenomena. The structural content of RDCs will be briefly explained. Finally some model will be presented to describe how the dynamic information is encoded, and extracted, from RDCs.

1 Considering protons, observed at 14.1T, in exchange between two sites different by 1ppm the corresponding timescale in roughly one millisecond.

2.2 EXPERIMENTAL ASPECTS OF RESIDUAL DIPOLAR COUPLING MEASUREMENTS

2.2.1 *Partial Orientational Order Effects in NMR Spectra*

The existence of orientational order in the sample will modify the appearance of the spectrum as the different anisotropic interactions will no longer be averaged to their isotropic values. Chemical shifts, dipolar and quadrupolar couplings are thus modulated by this incomplete averaging integral.

CHEMICAL SHIFT. The anisotropy of the chemical shift tensor moves the frequency to the orientationally averaged value [15]. The information content of chemical shift anisotropy (CSA) is in principle similar to that of RDCs as it is modulated by the averaged orientation of the CSA tensor and is sensitive to a similar timescale window. Nevertheless, for proteins, shifts induced by CSA are often very small, less than 0.1 ppm, and can be biased by the fact that changing the medium, by adding an alignment reactant, can bias chemical shift positions. Characterization of structural information of CSA has been proposed by many authors [54–56] but fewer applications have exploited residual CSA to understand slow dynamic averaging [57]. CSA is of course of great importance concerning fast dynamics as it consitutes a relaxation active interaction (see more detailed characterization in Chapter 1).

DIPOLAR COUPLING. The dipolar interaction in solution state NMR results in a splitting, or an additional modulation of the splitting observed between two scalar coupled spins. Therefore the dipolar splitting between two scalar coupled spins will be the sum of the J coupling and the RDC[2]. The absolute value of the splitting can both increase or decrease depending on the sign of the coupling. Typical values of RDC between N_i-H_i^N amide bounds ($^1D_{NH}$) lie in the range of $\pm$20 Hz and are obtained by measuring the difference between the observed splitting in the aligned sample compared to the splitting measured in the non-aligned sample.

Increasing the alignment will increase the range of the coupling, which intrinsically increases the accuracy of the measurement. However, this will also increase the influence of the so-called stray-couplings, especially long-range couplings between protons, and therefore dramatically increase spectral complexity in large spin systems such as proteins. Moreover, an excessive increase of order will bring the system to a more complex situation

2 The J coupling can exhibit anisotropic contribution too, but this effect is often indistinguishable from the RDC contribution and much smaller, it is therefore completely neglected [15].

where spins are strongly coupled [15]. This will lead to a clear degradation of spectral qualities.

QUADRUPOLAR COUPLING. In some cases the quadrupolar interaction can appear in partially ordered samples [15]. This interaction occurs only for nuclei of spin greater than 1/2 and corresponds to the interaction of the electric quadrupole moment of the nucleus with the surrounding electric gradient. It will usually not have any effect on protein spectra due to their nucleide composition, but will impact deuterium ($S = 1$) spectra. This effect will split the single HOD resonance observed in an isotropic system in partially deuterated water into a doublet where the spacing can be used, for a given alignment medium, as a rough estimation of the level of alignment of the sample.

2.2.2 *The Different Kinds of Alignment Media*

As previously mentioned the alignment of a molecule will be possible only if orientational order is present in the system. The way by which the medium transfers its orientational order to the solute may not be obvious, but it has to be mediated through weak interactions. A key question concerns whether the structural and dynamic properties of the solute remain unchanged in the presence of the alignment medium [48]. This can be qualitatively verified through the observation of localised shifts in the resonance frequency of certain peaks, although these effects might actually be expected to be very small if the alignment is weak. In terms of physical interactions, four different sources of alignment are known through magnetic, positive and negative electrostatic or steric interactions.

Different properties are required to obtain a useful alignment medium: it must be chemically inert, stable over time and under a large range of chemical environments, easy to manipulate, not too expensive and capable of inducing a sufficiently small level of alignment (on the order of one in a thousand). In response to these requirements a large range of media have been developed in the last two decades [19, 48].

MAGNETIC ALIGNMENT. This kind of alignment occurs for any compounds with an anisotropic magnetic susceptibility [58]. Biomolecular systems that spontaneously align to a sufficient degree to obtain measurable RDCs are mainly metallo-proteins containing a paramagnetic ion [52] or nucleic acids [59, 60] where the anisotropic susceptibility of the bases can constructively interact. If the system does not exhibit large enough anisotropic magnetic susceptibly, it has been proposed to attach a tag containing paramagnetic metals, which would provide the source of alignment [61, 62].

This kind of alignment, if due to a paramagnetic center with an anisotropic magnetic susceptibility tensor, will also lead to the appearance of pseudo contact shifts that are the manifestation of the dipolar interaction between the nucleus magnetic moment and the one generated by the differences in population of the various electron spin states [58].

LIQUID CRYSTALS. The diversity of dilute liquid crystals proposed to align biological samples is nowadays quite vast and the proposed alignment systems behave principally as nematic or smectic phases. Schematically, nematic phases, named from the greek word for thread, are rod shaped and usually exhibit ordered behavior in a single direction and smectic phase, from the greek word soap, are more ordered than the nematic as they present layered structures [47] (see Figure 3). Most are lyotrophic which means that their phase diagram is mainly dependent on two parameters, the temperature (as for thermotrophic ones) and the concentration of the chemicals that will build the structure into the solvent.

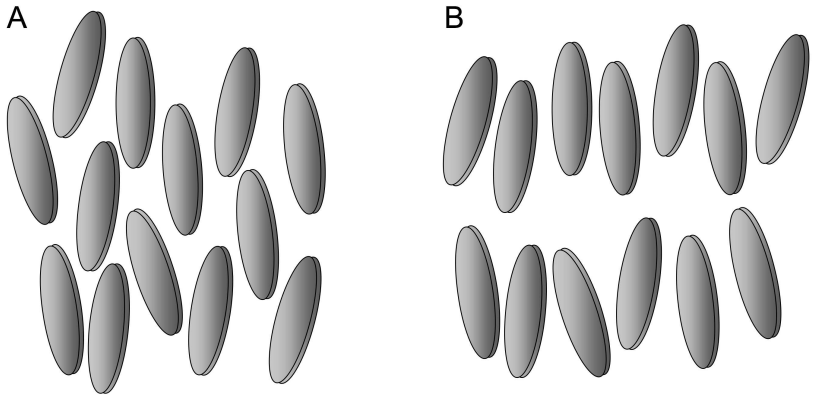

Figure 3 – Schematic representation of nematic and smectic phases. Often nematic liquid crystal (A) exhibit order in a single direction, whereas smectic phases (B) present in addition a layered structure.

The first dilute liquid crystal used for aligning proteins was made of a mixture of two phospholipid [53]: one long the DMPC (dimyristoyl-phosphatidylcholine) and one short the DHPC (dihexanoyl- phosphatidyl-choline) which form disks with a diameter around 400 Å. They are called bicelles and their surface is constituted mainly of the long alkyl chain phospholipid and the border mainly of the short chains [63, 64]. Charged groups point into the water whereas the aliphatic chains constitute the bulk of the bicelle. The alignment, that occurs at temperature around 35 °C or above, is induced by the diamagnetic susceptibility of the bicelles and their director is orthogonal to the magnetic field [64, 65]. They mainly interact with solutes through steric interactions. Different compositions [66] have now been

proposed to extend the use of bicelles over a wider range of pH [67, 68], temperature [67, 69, 70] or to increase their time stability [68, 71].

Bicelles can also be electrostatically charged [68, 71, 72] using CTAB (hexadecyl-trimethyl-ammonium bromide) or SDS (sodium dodecyl-sulfate) in order to switch to positive or negative electrostatic interactions with the solute or doped with lanthanides [69, 73] that will flip the director parallel to the magnetic field.

Other common alignment media can be found in filamentous bacteriophage such as fd and tobacco mosaic virus (TMV) [74] or more often the Pf1 [75, 76] that presents an elongated rod shape of ~20 000 Å long and a diameter of ~60 Å. They are made of DNA (deoxyribonucleic acid) encapsulated by negatively charged α-helical proteins and align through their intrinsic magnetic susceptibility apparently due to a repetitive distribution of planar peptide bonds in the protein shell [77]. The interaction with a solute will mainly be due to negatively charged electrostatic interactions. The electrostatic interaction can be reduced by increasing the ionic strength of the solution [78] thereby avoiding attractive interaction, and binding, with positively charged solutes.

The mixture of *n*-alkyl-poly(ethylene glycol) (PEG) and *n*-hexanol provides an other source of alignment [79] forming lyotropic lamellar phases. Layers are organized with hydrophobic parts inside, and PEG polar head-group in contact with water. They align parallel to the magnetic field and the spacing between layers can be tuned from several hundred to a few nanometers by increasing the concentration. Variations using different alkyl chains or by substituting the PEG with cetylpyridinium chloride or bromide have been proposed [79–81]. PEG-hexanol mixtures are cheap, while the large accessible range of pH and temperature makes them particularly attractive for studying a broad range of proteins.

Other liquid crystal based alignment media have been proposed, such as negatively charged purple membrane fragments containing bacteriorhodopsin [82, 83], cellulose micro-crystallites [84–86] that form rods of approximately 2000 Å long and 100 Å wide, that align orthogonally to the magnetic field, stacked columns of d(GpG) (2′-deoxyguanylyl-(3′,5′)-2′-deoxyguanosine) [87] which are compatible with detergents or collagen [88].

GEOMETRICALLY STRAINED SYSTEMS. A common form of alignment that is used for biomolecular systems results from mechanically straining polyacrylamide gels [89, 90]. They can be obtained for example by compressing a gel in a Shigemi tube by pressure of the plunger (compressed gels) or by re-swelling a dried gel in an NMR tube whose diameter is smaller than the

one where the gel was polymerized (stretched gels). This mechanical strain creates anisotropic cavities in the gel. The alignment is transfered to the solute through orientational restriction: if the molecule has for example an ellipsoidal shape, the main axis will have a higher probability to stand in the longest direction of the cavity than in the shortest one.

Gels can be charged during polymerization by including positively or negatively charged or zwitterionic monomers instead of neutral ones [91, 92]. This will modify the alignment properties of the medium by introducing electrostatic interactions. Gels can also be polymerized in the presence of phages [93, 94] or purple membranes [90] for example in order to give composite alignment media.

This section explained experimental manifestations induced by using anisotropic media, the following parts will focus on the dipolar interaction, the source of those observed phenomena.

2.3 THE DIPOLAR INTERACTION

2.3.1 *Origin of the Dipolar Interaction*

Each magnetic moment $\boldsymbol{\mu}_i$, such as the one induced by the spin of a nucleus, generates a local magnetic field $\mathbf{B}_i$. Therefore two moments $\boldsymbol{\mu}_I$, $\boldsymbol{\mu}_S$ will interact through this field leading to the following interaction energy [17, 95]:

$$E_{IS} = -\boldsymbol{\mu}_I \cdot \mathbf{B}_S = -\boldsymbol{\mu}_S \cdot \mathbf{B}_I \tag{2.1}$$

with

$$\mathbf{B}_i\left(\mathbf{r}_{IS}\right) = \frac{\mu_0}{4\pi}\frac{1}{r_{IS}^3}\left[3\left(\boldsymbol{\mu}_i \cdot \boldsymbol{e}_{IS}\right)\boldsymbol{e}_{IS} - \boldsymbol{\mu}_i\right] \tag{2.2}$$

where $\mathbf{r}_{IS}$ is the vector connecting the two particles carrying the magnetic moments and $\boldsymbol{e}_{IS} = \mathbf{r}/r_{IS}$ the associated unitary vector, as shown in Figure 4.

Considering two nuclear spins I and S, their associated magnetic moments operators are:

$$\hat{\boldsymbol{\mu}}_I = \gamma_I \hbar \hat{\mathbf{I}} \qquad \text{and} \qquad \hat{\boldsymbol{\mu}}_S = \gamma_S \hbar \hat{\mathbf{S}} \tag{2.3}$$

and using the correspondence principle [14], we obtain the following Hamiltonian for the dipolar interaction:

$$\mathcal{H}_{IS}^{D} = d_{IS}\left[3\left(\hat{\mathbf{I}} \cdot \boldsymbol{e}_{IS}\right)\left(\hat{\mathbf{S}} \cdot \boldsymbol{e}_{IS}\right) - \hat{\mathbf{I}} \cdot \hat{\mathbf{S}}\right] \tag{2.4}$$

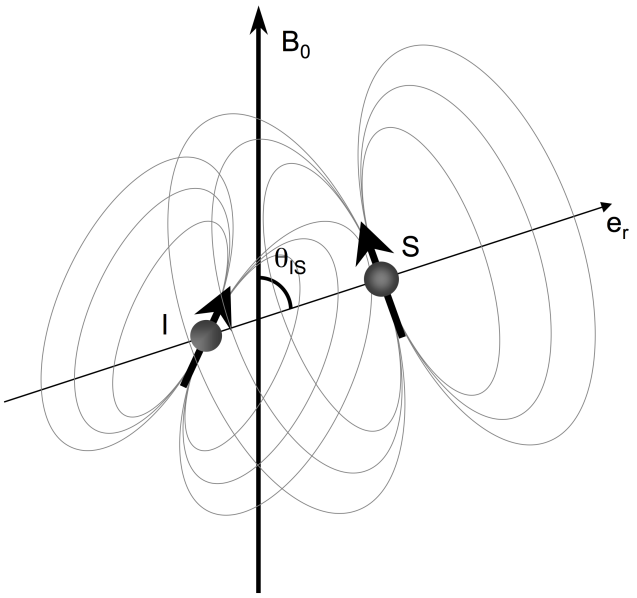

Figure 4 – Magnetic fields generated by the two magnetic moments of spin I and S, source of the dipolar interaction and representation of θ_{IS} the angle between the internuclear vector and the $\mathbf{B}_0$ field.

where, if the Hamiltonian is expressed in frequency units:

$$d_{IS} = -\frac{\gamma_I \gamma_S \mu_0 h}{16\pi^3 r_{IS}^3} \tag{2.5}$$

2.3.2 *High Field Approximation and Weak Coupling Limit*

In solution NMR, the preponderant interaction that gives the overall feature of the NMR spectrum is the Zeeman interaction. The magnitude of this interaction is proportional to the magnetic field strength and the field direction will, define the z-axis direction of the laboratory frame used in this study.

The high field approximation consists of considering the dipolar interaction as a perturbation of the Zeeman Hamiltonian $\mathcal{H}^Z$. Using first order perturbation development [95] only the part of the interaction that commute with $\mathcal{H}^Z$, which are called secular terms, will remain. Thus for an energy transition only phenomena occurring at similar frequencies will generate non-negligible probabilities of transition and the Zeeman energy levels will be modified according to the simplified Hamiltonian [15–17, 95]:

$$\mathcal{H}_{IS}^D(\theta_{IS}) = d_{IS}\frac{3\cos^2\theta_{IS} - 1}{2}\left(3\hat{I}_z\hat{S}_z - \hat{\mathbf{I}}\cdot\hat{\mathbf{S}}\right) \tag{2.6}$$

where θ_{IS} is the angle between the internuclear vector and the $\mathbf{B}_0$ field (illustrated in Figure 4).

This expression can be further simplified in the case of weak coupling [15, 16, 95] where the energy difference between each Zeeman energy level is much bigger than the dipolar coupling interaction. Here all non diagonal terms of the dipolar interaction are neglected whereas within the secular approximation a block-diagonal form can remain. The Hamiltonian expression becomes:

$$\mathcal{H}_{IS}^{D}(\theta_{IS}) = d_{IS}\frac{3\cos^2\theta_{IS} - 1}{2} 2\hat{I}_z\hat{S}_z \tag{2.7}$$

The approximation is valid if the sum of the couplings (scalar J_{IS} and dipolar D_{IS}) between the two spins I and S is much smaller than the chemical shift difference $\Delta\omega_{IS}$ [15], that is to say:

$$|\Delta\omega_{IS}| \gg \frac{1}{2}|D_{IS} + J_{IS}| \tag{2.8}$$

Due to the large differences in gyromagnetic ratio between nuclei, and therefore between their Larmor frequencies, this approximation is always valid for heteronuclear spin systems [16]. Concerning homonuclear systems, this hypothesis will be considered as valid in the following study and can be qualitatively justified: as mentioned in Section 2.2 RDCs are often less than 10 Hz which roughly corresponds to a chemical shift difference of 0.02 ppm for proton or 0.1 ppm for carbon at 14.1 T $\mathbf{B}_0$ field and J-couplings are often less than 20 Hz for proton couplings and less than 100 Hz for carbons, leading to total couplings smaller than 0.05 ppm for protons and 1 ppm for carbon, which considering standard $|\Delta\omega_{IS}|$ usually verify expression 2.8.

In the weak coupling interaction limit the magnitude of the Hamiltonian governing the dipolar interaction in the IS spin system (see equation 2.7) which has a similar form to the scalar coupling Hamiltonian [15, 16]. Therefore the measurable manifestation of the dipolar interaction, will appear through an additional contribution to the apparent J coupling.

From now this approximation will be considered, but its validity will of course depend on the amount of orientational order present in the system. The following section will present how this degree of order modulates the averaging of dipolar couplings in solution in order to derive analytical expressions for RDCs.

2.4 THE DIPOLAR INTERACTION AVERAGING IN LIQUID STATE NMR

2.4.1 *Motional Averaging*

In the weak coupling approximation the expression of the dipolar Hamiltonian is modulated by $d_{IS}P_2(\cos\theta_{IS})$ where P_2 is the second order Legendre polynomial and therefore depends on the orientation of the internuclear vector of interest with respect to the $\mathbf{B}_0$ field.

During a measurement, in liquid state NMR, this vector will be modified as soon as some dynamic phenomenon occurs. This evolution is both time and ensemble averaged but supposing an ergodic behavior for this system it will be assumed that the two kinds of averaging lead to identical results. All dynamics with characteristic time scales faster than the chemical shift coalescence limit can influence this averaging process e.g. molecular reorientation, domain motion, local dynamics... Experimentally obtained RDCs can be theoretically calculated by averaging the term $d_{IS}P_2(\cos\theta_{IS})$ [15].

Considering covalently bound nuclei, the magnitude dependence present in d_{IS} through $1/r_{IS}^3$ is the easiest to treat in the sense that it occurs on very fast timescales [15, 96], picoseconds or faster, which is much faster than other molecular motions. This confers a statistical independence [46] and this motion, sometimes called libration, can be taken into account in the d_{IS} expression just by replacing the instantaneous internuclear distance by an effective distance [97]. To avoid over-complicating the notation this average will not be explicitly shown in all formulae but will be from now always implicitly assumed.

For the angular dependence one can assume that local dynamics and global reorientation occurs in an uncorrelated manner and subsequent averages can therefore be applied successively.

Starting with the global reorientation two cases can be distinguished:

- the isotropic case, where the probability to find the internuclear vector oriented with an angle θ towards the $\mathbf{B}_0$ field is given by:

$$p(\theta) = \frac{\sin\theta}{\int_0^\pi \sin\theta d\theta} = \frac{\sin\theta}{2} \tag{2.9}$$

which leads to a null average for $P_2(\cos\theta)$ as:

$$\begin{aligned}\left\langle P_2(\cos\theta)\right\rangle_{iso} &= \frac{1}{2}\int_0^{\pi}\frac{3\cos^2\theta-1}{2}\sin\theta d\theta \\ &= -\frac{3}{4}\int_0^{\pi}\cos^2\theta d(\cos\theta)+\frac{1}{4}\int_0^{\pi}\sin\theta d\theta \\ &= 0\end{aligned} \tag{2.10}$$

- and the anisotropic case, where all orientations are no longer equiprobable and therefore a non zero-coupling can be expected.

As a consequence, dipolar couplings will be directly unobservable in isotropic liquids, whereas in anisotropic systems some residual dipolar couplings will remain. If $\langle . \rangle$ represents the average over all different kinds of motion, the RDC can be expressed as follows:

$$D_{IS} = \left\langle d_{IS} P_2\cos(\theta_{IS})\right\rangle \tag{2.11}$$

which as previously explained simplifies for covalently bound vectors, if an effective internuclear distance is introduced, as [19]:

$$D_{IS} = d_{IS}\left\langle P_2\cos(\theta_{IS})\right\rangle \tag{2.12}$$

2.4.2 *Incomplete Averaging of the Dipolar Interaction in Anisotropic Liquids*

In order to study the orientational distribution of a given vector it is convenient to introduce a frame bound to the protein structure called the molecular frame. In this way it becomes possible to express the global properties of alignment of the system by characterizing the orientation of the $\mathbf{B}_0$ field in this frame and to study the local motion, in a more intuitive way, in this local frame. Using director cosine formalism, as shown in Figure 5, the $\cos\theta_{IS}$ term can be developed in:

$$\begin{aligned}\cos\theta_{IS} = \boldsymbol{e}_{IS}\cdot\boldsymbol{e}_{B_0} &= \begin{pmatrix}\cos\alpha_x\\ \cos\alpha_y\\ \cos\alpha_z\end{pmatrix}\cdot\begin{pmatrix}\cos\beta_x\\ \cos\beta_y\\ \cos\beta_z\end{pmatrix} \\ &= \cos\alpha_x\cos\beta_x+\cos\alpha_y\cos\beta_y+\cos\alpha_z\cos\beta_z\end{aligned} \tag{2.13}$$

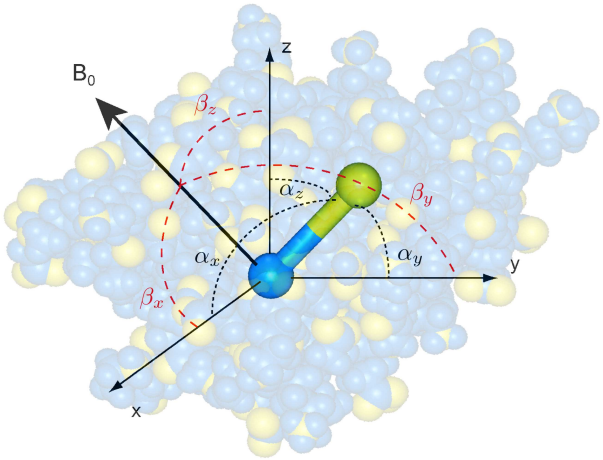

Figure 5 – Orientation of the $\mathbf{B}_0$ field and the internuclear vector in the molecular frame bound to the protein structure $(\boldsymbol{e}_x, \boldsymbol{e}_y, \boldsymbol{e}_z)$ are described using $(\beta_x, \beta_y, \beta_z)$ and $(\alpha_x, \alpha_y, \alpha_z)$ respectively.

which give for the corresponding second order dynamically averaged Legendre polynomial of the dipolar interaction:

$$\left\langle \frac{3\cos^2\theta - 1}{2} \right\rangle = \frac{3}{2}\left\langle \left(\cos\alpha_x \cos\beta_x + \cos\alpha_y \cos\beta_y + \cos\alpha_z \cos\beta_z\right)^2 \right\rangle - \frac{1}{2} \tag{2.14}$$

Assuming that overall reorientation and local motion are statistically independent, it become possible to average the two motions independently and the previous expression becomes:

$$\left\langle \frac{3\cos^2\theta - 1}{2} \right\rangle = \frac{3}{2} \sum_{k,l\in\{x,y,z\}} \left\langle \cos\beta_k \cos\beta_l \right\rangle_O \left\langle \cos\alpha_k \cos\alpha_l \right\rangle_I - \frac{1}{2} \tag{2.15}$$

where $\langle . \rangle_O$ and $\langle . \rangle_I$ represent dynamical averages through overall and internal motion respectively.

To describe the properties of overall orientational averaging or alignment of the molecule we can introduce a unit-less symmetric order matrix $\hat{\mathbf{A}}$, called the Saupe matrix that characterizes alignment in the form of a traceless second rank tensor, defined as [98]:

$$A_{kl} = \frac{1}{2}\left(3\left\langle \cos\beta_k \cos\beta_l \right\rangle_O - \delta_{kl}\right) \qquad k, l \in \{x, y, z\} \tag{2.16}$$

where δ_{kl} the Kronecker symbol. This allows to rewrite equation 2.12 as:

$$D_{IS} = d_{IS} \sum_{k,l} A_{kl} \left\langle \cos\alpha_k \cos\alpha_l \right\rangle_I \tag{2.17}$$

It is worth noting that the Saupe matrix is by construction traceless, real and symmetric. The two last properties ensure that this matrix can be diagonalised and the associated eigenbasis, whose eigenvectors are ordered by convention according to $|A_{xx}| \leqslant |A_{yy}| \leqslant |A_{zz}|$ and normalized for convenience, will be called the Principal Axis System (PAS). Switching from the molecular frame to the PAS can be achieved through the three dimensional Euler rotations, where three successive rotations of angle α, β, γ are used to build the matrix $R(\alpha, \beta, \gamma)$ describing the rotation between the two frames. In the PAS frame the previous equation becomes:

$$D_{IS} = d_{IS}\left(A_{xx}\left\langle\cos^2\alpha_x\right\rangle_I + A_{yy}\left\langle\cos^2\alpha_y\right\rangle_I + A_{zz}\left\langle\cos^2\alpha_z\right\rangle_I\right) \tag{2.18}$$

which gives, using spherical coordinates (θ, ϕ):

$$D_{IS}(\theta,\phi) = d_{IS}\Big(A_{xx}\left\langle\sin^2\theta\cos^2\phi\right\rangle_I + A_{yy}\left\langle\sin^2\theta\sin^2\phi\right\rangle_I + A_{zz}\left\langle\cos^2\theta\right\rangle_I\Big) \tag{2.19}$$

If axial A_a and rhombic A_r components are introduced as:

$$A_a = \frac{A_{zz}}{2} \qquad \text{and} \qquad A_r = \frac{A_{xx} - A_{yy}}{3} \tag{2.20}$$

the expression 2.19 can be rewritten as:

$$D_{IS}(\theta,\phi) = d_{IS}\left[A_a\left\langle 3\cos^2\theta - 1\right\rangle_I + \frac{3}{2}A_r\left\langle\sin^2\theta\cos 2\phi\right\rangle_I\right] \tag{2.21}$$

Or using an equivalent formalism where the rhombicity, $R = A_r/A_a$ is used, in order to quantify the "shape" of the alignment tensor in a magnitude independent way:

$$D_{IS}(\theta,\phi) = d_{IS}A_a\left[\left\langle 3\cos^2\theta - 1\right\rangle_I + \frac{3}{2}R\left\langle\sin^2\theta\cos 2\phi\right\rangle_I\right] \tag{2.22}$$

This description of the tensor will be often preferred to the Saupe matrix description. The approaches are perfectly equivalent (one is the diagonal form of the other) and both require five parameters to be completely described: the three upper or lower off diagonal elements and two diagonal terms for the Saupe matrix and three Euler angles, a magnitude and a rhombic component for the molecular alignment tensor description.

By introducing the Saupe matrix it has been implicitly assumed that local dynamics does not affect the properties of alignment of the molecule, that is to say that the system can be described with a single order matrix. This approximation may be inappropriate if the system undergoes large conformational changes at a timescale faster than the millisecond for example.

In this case it is possible to use such a description by introducing a set of molecular conformations that are representative of this large configurational modification and which are all described by their own order matrix, and then estimate the expected RDC by taking the averaged over all these conformers (for more details see Section 2.6).

The use of the Euler formalism is especially practical for transforming spherical harmonics $Y_{l,m}(\theta,\phi)$ from one frame to another. Thus it is useful to rewrite equation 2.21 in terms of the spherical harmonics:

$$D_{IS} = d_{IS}\sqrt{\frac{16\pi}{5}}\left[A_a\left\langle Y_{2,0}(\theta,\phi)\right\rangle_I + \sqrt{\frac{3}{8}}A_r\left(\left\langle Y_{2,-2}(\theta,\phi)\right\rangle_I + \left\langle Y_{2,2}(\theta,\phi)\right\rangle_I\right)\right] \tag{2.23}$$

where the rank 2 spherical harmonics are defined as [14]:

$$\begin{aligned} Y_{2,0}(\theta,\phi) &= \sqrt{\frac{5}{16\pi}}(3\cos^2\theta - 1) \\ Y_{2,\pm1}(\theta,\phi) &= \mp\sqrt{\frac{15}{8\pi}}\sin\theta\cos\theta\, e^{\pm i\phi} \\ Y_{2,\pm2}(\theta,\phi) &= \sqrt{\frac{15}{32\pi}}\sin^2\theta\, e^{\pm 2i\phi} \end{aligned} \tag{2.24}$$

If we now consider two frames $\mathcal{R}$ and $\mathcal{R}'$ where the orientation of the vector of interest is given respectively by (θ,ϕ) and (θ',ϕ') and if the passage from $\mathcal{R}$ to $\mathcal{R}'$ is defined by the three Euler angles (α,β,γ), it becomes possible to express any spherical harmonics in $\mathcal{R}$ as a linear combination of the spherical harmonics of the same order in $\mathcal{R}'$. Considering rank 2 spherical harmonics (the only ones necessary for RDCs) this linear combination can be expressed as:

$$Y_{2,p}(\theta,\phi) = \sum_{q=-2}^{2} D^{(2)}_{q,p}(\alpha,\beta,\gamma)Y_{2,q}(\theta',\phi') \tag{2.25}$$

where $D^{(2)}_{q,p}$ are rank two Wigner matrices, which can be expressed using reduced Wigner matrices $d^{(2)}_{q,p}$ as:

$$D^{(2)}_{q,p}(\alpha,\beta,\gamma) = e^{-iq\alpha}d^{(2)}_{q,p}(\beta)e^{-ip\gamma} \tag{2.26}$$

where reduced Wagner matrices can be obtained using the following expressions [99]:

$$\begin{aligned} d^{(2)}_{2,\pm2}(\beta) &= \left(\frac{1\pm\cos\beta}{2}\right)^2 & d^{(2)}_{1,\pm1}(\beta) &= \frac{1\pm\cos\beta}{2}(2\cos\beta\mp1) \\ d^{(2)}_{2,\pm1}(\beta) &= -\frac{1\pm\cos\beta}{2}\sin\beta & d^{(2)}_{1,0}(\beta) &= -\sqrt{\frac{3}{2}}\sin\beta\cos\beta \\ d^{(2)}_{2,0}(\beta) &= \frac{\sqrt{6}}{4}\sin^2\beta & d^{(2)}_{0,0}(\beta) &= \frac{3\cos^2\beta-1}{2} \end{aligned} \tag{2.27}$$

Using this formalism, a given RDC can be expressed in any frame. As expression 2.23 is valid in the PAS, here $\mathcal{R}$, it is possible to express it in an arbitrary molecular frame $\mathcal{R}'$ by the following expression:

$$\begin{aligned} D_{IS}(\theta,\phi) = d_{IS}\sqrt{\frac{16\pi}{5}}\Bigg[& A_a \sum_{q=-2}^{2} e^{-iq\alpha} d^{(2)}_{q,0}(\beta) \left\langle Y_{2,q}(\theta',\phi') \right\rangle_I \\ & + \sqrt{\frac{3}{8}} A_r \Bigg(\sum_{q=-2}^{2} e^{-iq\alpha} d^{(2)}_{q,-2}(\beta) e^{2i\gamma} \left\langle Y_{2,q}(\theta',\phi') \right\rangle_I \\ & + \sum_{q=-2}^{2} e^{-iq\alpha} d^{(2)}_{q,2}(\beta) e^{-2i\gamma} \left\langle Y_{2,q}(\theta',\phi') \right\rangle_I \Bigg) \Bigg] \end{aligned} \tag{2.28}$$

where (α, β, γ) are the Euler angles defining the rotation from the PAS to the local frame and where (θ', ϕ') reflect the vector orientation in this arbitrary molecular frame $\mathcal{R}'$.

Using spherical harmonics, it becomes possible to introduce an order parameter:

$$S^2 = \frac{4\pi}{5} \sum_{p=-2}^{2} \left\langle Y^*_{2,p}(\theta',\phi') \right\rangle \left\langle Y_{2,p}(\theta',\phi') \right\rangle \tag{2.29}$$

This can be compared to the order parameter commonly used to model the reorientational auto-correlation function of dipolar interactions in the analysis of spin relaxation data, where it corresponds to the plateau value of the internal correlation function.

2.5 STRUCTURAL INFORMATION FROM RESIDUAL DIPOLAR COUPLINGS

2.5.1 *The Static Approximation*

In the case of folded, stable structures, showing small amplitude dynamic fluctuations in terms of bond vector position and orientation, RDCs are first and foremost exquisitely sensitive probes of molecular conformation. Historically RDCs have therefore been exploited to refine and even determine three dimensional protein structures without explicitly accounting for local motion.

Under the assumption that the internuclear orientations and positions are fixed relative to the molecular frame the analytical expression of RDCs, given by equation 2.21 reduces to:

$$D_{IS}^{stat}(\theta,\phi) = d_{IS}\left[A_a\left(3\cos^2\theta - 1\right) + \frac{3}{2}A_r \sin^2\theta\cos 2\phi\right] \tag{2.30}$$

which can be expressed in cartesian coordinates as:

$$D_{IS}^{stat}(x,y,z) = d_{IS}\left[A_a\left(3\bar{z}^2 - 1\right) + \frac{3}{2}A_r(\bar{x}^2 - \bar{y}^2)\right] \tag{2.31}$$

where $\bar{x}$, $\bar{y}$, $\bar{z}$ are the normalized cartesian coordinates of the internuclear vector. This expression combined with the normalization restraint:

$$\bar{x}^2 + \bar{y}^2 + \bar{z}^2 = 1 \tag{2.32}$$

gives the ensemble of analytically acceptable solutions for an experimentally measured RDC. The expected value of an RDCs as a function of this orientation in the PAS is illustrated in Figure 6. Considering a single measurement

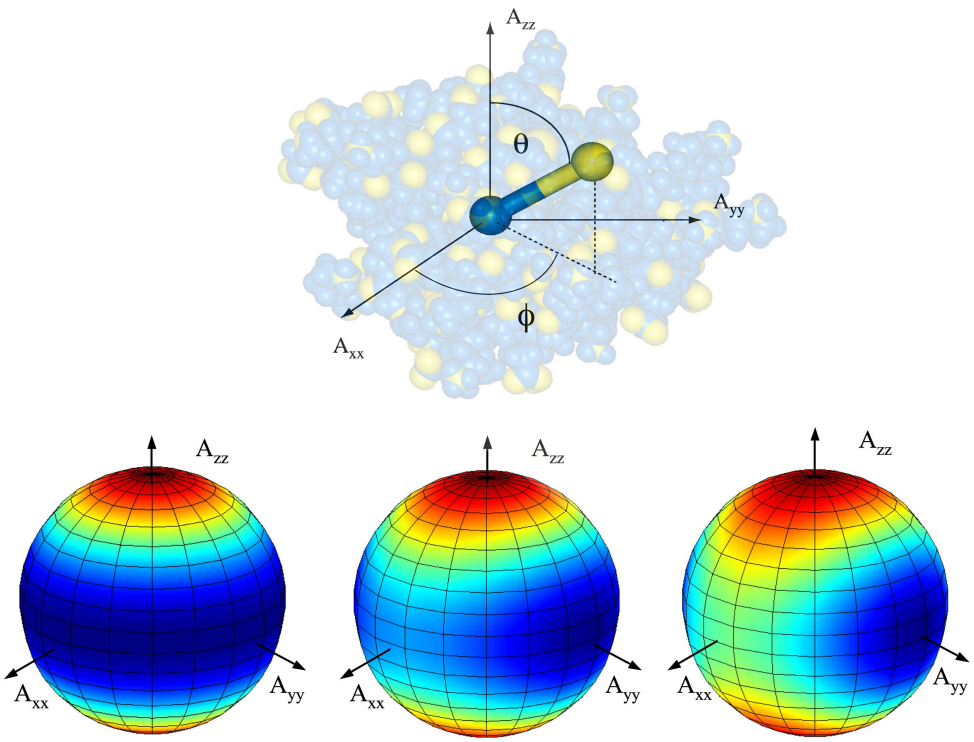

Figure 6 – Angular dependence of a static RDC in the PAS. Upper part: orientation of a internuclear vector in the PAS. Lower part: angular dependence of a static RDC. The value is encoded with a color scale. From dark red (higher value $D_{IS}^{max} = 2d_{IS}A_a$) to red, yellow, green, light blue and dark blue (lowest value $D_{IS}^{min} = -d_{IS}(1 + 3R/2)A_a$). From left to right the used tensor is not rhombic (left), or present increasing rhombicity (middle, right).

for a unique internuclear vector this exhibits more than one solution and the consequences of the orientational degeneracy will be now investigated.

2.5.2 *Orientational Degeneracy*

The orientational degeneracy will clearly depend on the studied system and the amount of experimental data available. Here the ensemble of solutions will be described for data measured in one alignment medium, for a single vector, a planar object or a chiral system.

SINGLE INTERNUCLEAR VECTOR. The ensemble of solutions for the system obeying equations 2.31 and 2.32 can be geometrically seen as the intersection of a quadric with the elementary sphere. The lowest degenerate solution presents a 2-fold symmetry which is obtained for the particular case where $\bar{z} = 1$ or $\bar{y} = 1$ and $R > 0$.

In the general case, the ensemble of solution is infinitely degenerate as it corresponds to two symmetrical unidimensional closed curve as shown in Figure 7.

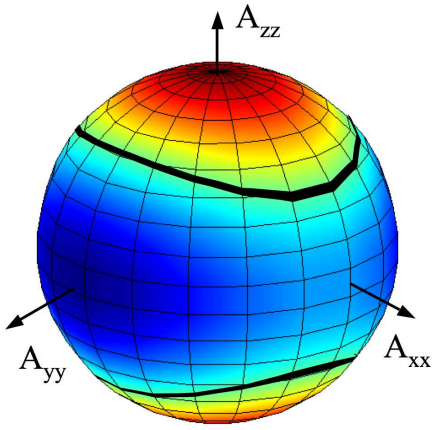

Figure 7 – Ensemble of possible orientations for an internuclear vector for which a single RDC is measured. Angular dependence of a static RDC in the PAS presented as a colored sphere. The black line indicate the ensemble of possible orientations, i.e. orientations that all give the value of a static RDC.

PLANAR SYSTEM. In order to decrease the large degeneracy of RDCs, it is possible to consider more complex systems of fixed geometry. For example, if a peptide plane is assumed to have a fixed planar topology in a protein it is possible to measure different RDCs in this system and therefore increase the amount of information available to orient this object. Switching to a two dimensional system, the ensemble of solutions has to respect 2.31 and 2.32 for any in-plane vector.

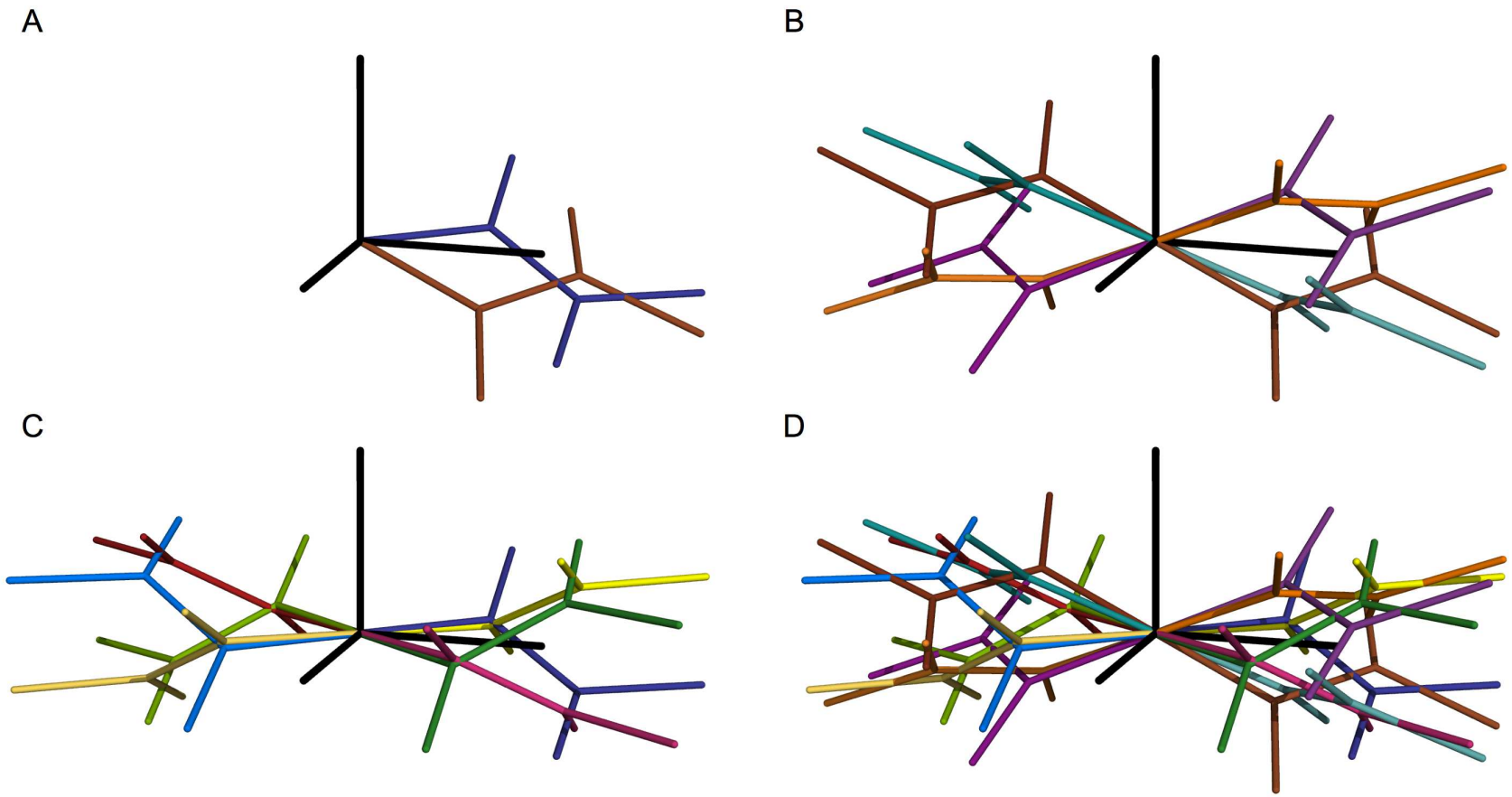

Figure 8 – Graphical representation of the 16-fold degeneracy for a peptide plane. Black frame represents the PAS. (A) Representation of a solution and its transform through the in plane transformation (B) and (C) 8-fold symmetry of those two solutions separately (D) all the 16 orientations. Atoms are not represented for clarity but are from the PAS origin and outwards: ${}^{\alpha}C_i$, N_i and H_i, C_{i-1} and O_{i-1} and ${}^{\alpha}C_{i-1}$.

It as been shown that the ensemble of solutions consists of at least 16 solutions [100]. Eight of the mathematical transformations that define the different solutions are clearly visible from equation 2.31, where the inversion around the three PAS axes ($x \to -x$, $y \to -y$, $z \to -z$) and their different combinations will leave the expected RDC unchanged. The eight others can be deduced from a symmetry action around an in-plane axis of the previously found solutions. The mathematical characterization of these solutions can be found in [100] and are graphically presented here in Figure 8.

CHIRAL OBJECT Different kinds of chiral systems can be found in proteins, from very small ones such as C^{α} tetrahedral junctions (for non-glycine natural amino-acid) to very large motifs such as a whole protein domain that is considered to be rigid. Chirality[3] will induce a decrease of the number of acceptable solutions compared to a planar system as it is a three dimensional object and therefore an even number of inversions will lead to an inversion of the chirality that is not physically acceptable. Consequently only a four fold degeneracy will remain, where all the solutions are related by two successive inversions around the PAS axis [19, 101], as illustrated in Figure 9.

2.5.3 *Multiple Alignment Media Information*

Considering any kind of structural motif, it is possible (in principle) to measure RDCs in different alignment media. If the alignment tensor changes, the degeneracy of any vector will be reduced as equation 2.31 has to be respected for all datasets. Of course the different alignment media have to be sufficiently independent in order to provide complementary information. For example changing the magnitude of the alignment or the direction of the director of the alignment system will not be enough to provide independent information [19].

SINGLE INTERNUCLEAR VECTOR. Switching from one to two alignment media drastically reduces the degeneracy of the solution ensemble. It will be reduced to eight orientations, however it often falls in practice to only four solutions. An example is shown in Figure 10, where the ensemble of solutions is presented as the intersection of the two independently determined ensembles of solutions of two different alignment media. If a third or even more media are added, the degeneracy will fall to two and the two solutions are related by complete inversion of their coordinates.

3 A chiral object is by definition non superposable to its mirror-image.

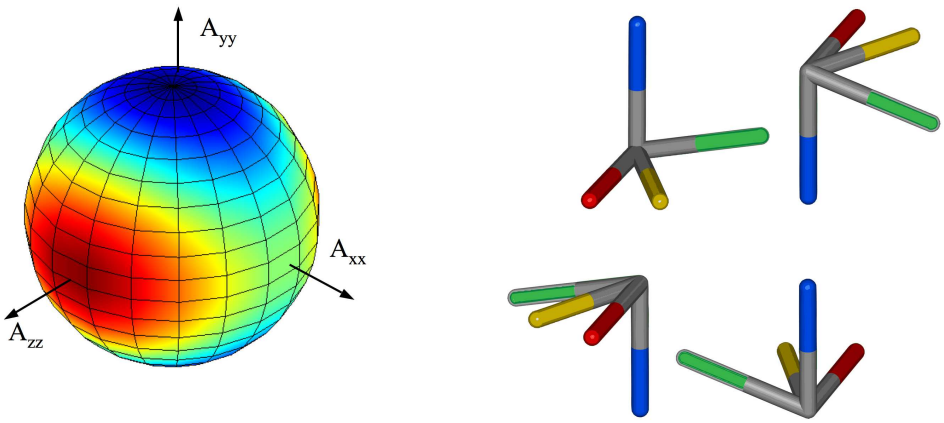

Figure 9 – Ensemble of acceptable solutions for a chiral object, for which RDCs are measured in a single alignment medium. The geometry of the chiral object is known and measured RDCs span at least three non collinear and non coplanar internuclear vector. A four fold degeneracy remains.

PLANAR SYSTEM. Adding a second (or even more) alignment medium will decrease the allowed solutions to two where the two solutions correspond to the intersection of the two ensembles of sixteen orientations acceptable by considering a unique alignment medium. The two solutions are mirror images of each other through a central symmetry and this degeneracy of two, intrinsic to the inherent information content of RDCs, cannot be raised by multiplying the number of alignment media.

CHIRAL OBJECT. Chirality is the only property that can completely overcome the orientational degeneracy. A single solution can be found as the intersection of the two ensembles satisfying RDCs for a single alignment medium (see Figure 11), but it can be easily deduced from the less restricted planar case where only two solutions exist, related by three consecutive planar inversions ($x \to -x$, $y \to -y$, $z \to -z$). For a chiral motif, it will lead to different chirality and therefore the solution as to be unique.

In conclusion, RDCs can be considered as an exquisitely sensitive source of structural information. Even if this information can be highly degenerate,

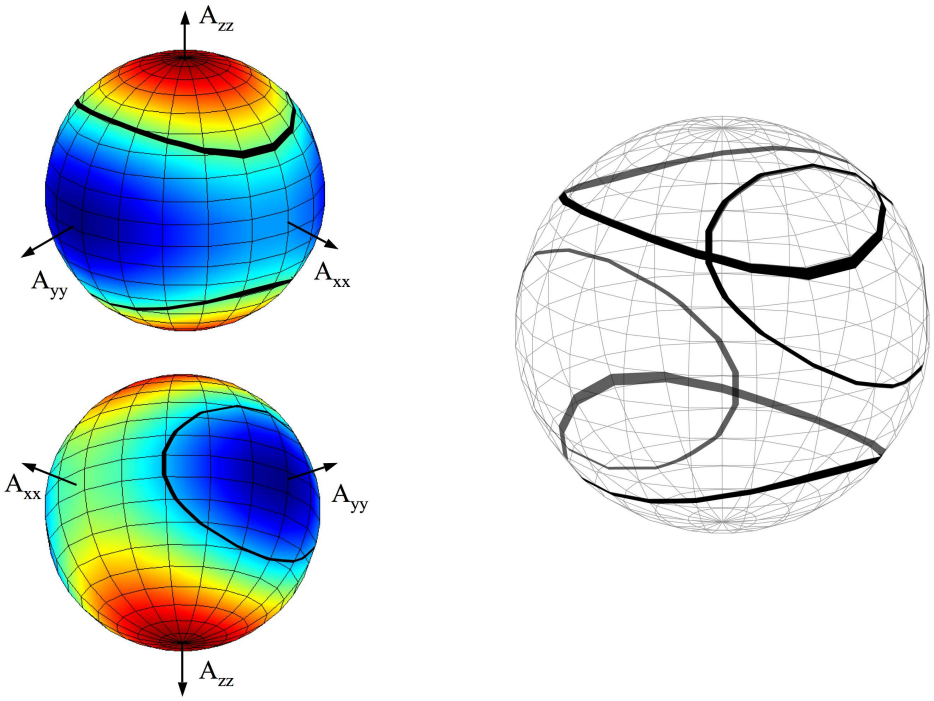

Figure 10 – Ensemble of solutions for an internuclear vector for which RDCs were measured in two alignment media. Left part: the ensemble of solutions (black line) for each medium considered independently. Right part: ensemble of solution arising from the intersection of the two single medium solution ensemble. In this particular case four solutions are possible. Two in the front part (back) and two on the back (grey part).

reasonable choice of structural motif, and — or — the use of several alignment media can drastically reduce this orientational ambiguity. Note that in the case of covalently bound nuclei this information is only orientational: it characterizes the orientation of a given vector in the PAS or in any molecular frame.

This information is highly complementary to that arising from ^{1}H-^{1}H nuclear Overhauser effect (nOe) which provides a network of distances between atoms used as restraints for structure calculation. This information is very efficient in providing the correct overall fold of a protein as only few long-range nOe can be sufficient to obtain a satisfying structure. Nevertheless the accuracy of these measurements results often in loose restraints and consequently to poor local structure definition [100]. RDCs can accurately and unambiguously define this local structure. It should also be mentioned that the resolution of orientational degeneracy for planar motifs connected at tetrahedral junctions has led to the determination of ultra-high resolution

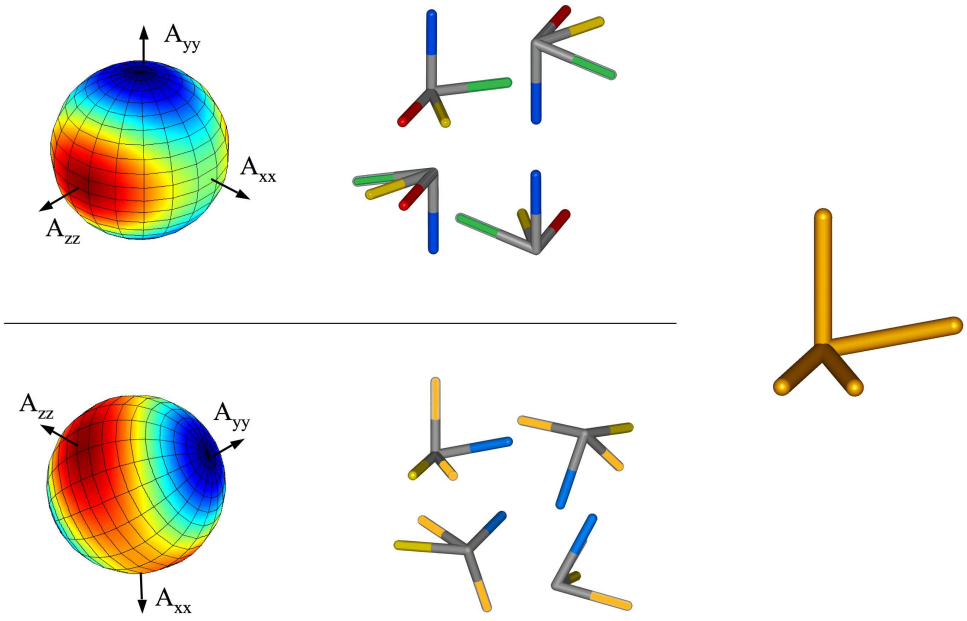

Figure 11 – Ensemble of acceptable solutions for a chiral object, for which RDCs are measured in two alignment media, presented as the intersection of the two four solution ensembles resulting from considering a single medium.

structures using only RDCs measured in two different alignment media [102].

2.6 DYNAMICAL MODELS FOR INTERPRETING RDCS

As underlined previously in this section, RDCs are clearly influenced by any kind of motion occurring at timescales up to the millisecond. Here, the effects of different models of local motion are discussed in the context of the extraction of dynamical information from experimental RDCs.

2.6.1 *Ensemble Averaging*

Although the static structural approach described above can in many cases define the structural characteristics of the protein with sufficient accuracy, dynamical characterization may require an ensemble average description:

$$D_{IS}^{ens} = \langle D_{IS,n} \rangle_{ens} = \sum_{n} p_n D_{IS,n} \tag{2.33}$$

where the sum is taken over n different conformers of the ensemble, each of them characterized by a statistical weight p_n[4]. This statistical weighting is for simplicity often approximated as a equiprobable distribution, but more sophisticated analysis can be carried out using Boltzmann weighting.

In order to simplify equation 2.33 different approximations can be made. Firstly if we consider covalently bound nuclei, the effective distance present in d_{IS} can be considered as identical for all conformers, which gives:

$$D_{IS}^{ens} = d_{IS} \sum_n p_n \left[A_{a,n} \left(3\cos^2\theta_n - 1 \right) + \frac{3}{2} A_{r,n} \sin^2\theta_n \cos 2\phi_n \right] \quad (2.34)$$

where $A_{a,n}$ and $A_{r,n}$ describe the alignment tensor of the n-th conformer and (θ_n, ϕ_n) the normalized spherical coordinates of the vector relative to this tensor.

It is worth noting that this expression allows to have a different alignment tensor for each of the conformers. It is thus possible to apply this description to very dynamic systems, such as unfolded proteins, where the internal flexibility is so large that local dynamics will clearly impact the properties of alignment [103].

If now one assumes that conformational fluctuation does not influence significantly the properties of alignment of the studied system, equation 2.34 further simplifies to:

$$D_{IS}^{ens}(\theta, \phi) = d_{IS} \sum_n p_n \left[A_a \left(3\cos^2\theta_n - 1 \right) + \frac{3}{2} A_r \sin^2\theta_n \cos 2\phi_n \right] \quad (2.35)$$

or if $\langle . \rangle_n$ represents the ensemble average we obtain:

$$D_{IS}^{ens} = d_{IS} \left[A_a \left\langle 3\cos^2\theta - 1 \right\rangle_n + \frac{3}{2} A_r \left\langle \sin^2\theta \cos 2\phi \right\rangle_n \right] \quad (2.36)$$

which exactly corresponds to equation 2.21, where the internal dynamics has been expressed as an ensemble average.

If a structural ensemble is known and a single alignment tensor postulate, a convenient way to obtain relevant alignment tensors is to use Singular Value Decomposition (SVD) as explained in Annexe A.

This model is the basis of most of the numerical approaches to describe dynamics from experimental RDCs. Nevertheless analytical descriptions can also be useful to understand the details of physical phenomena and we will now focus on such models.

4 For this particular expression (2.33) the use of an effective distance is not necessary: the point it especially interesting for characterizing long-range RDCs.

2.6.2 *Axially Symmetric Motion*

An axially symmetric motion commonly used in NMR is the diffusion in a cone model. This model describe the dynamics of a vector through a circular cone of opening angle 2Θ about the average orientation of the vector (θ_{av}, ϕ_{av}). The vector is allowed to freely diffuse within this volume. The study of this motion (illustrated in Figure 12) will be described in a local frame whose z'-axis is along the height of the cone.

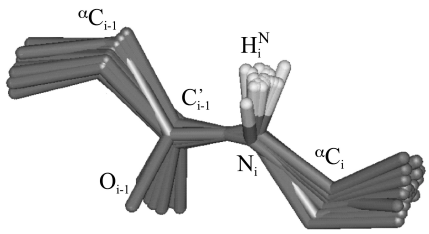

Figure 12 – Representation of the diffusion in a cone motion for a peptide plane.

As the diffusion occurs in an uncorrelated manner between the zenith and azimuth angle, they can be independently averaged. Due to the symmetry of the motion, the azimuthal dependent part of the spherical harmonics (see equation 2.24) averages to zero:

$$\left\langle e^{\pm i\phi'} \right\rangle_{axial} = \left\langle e^{\pm 2i\phi'} \right\rangle_{axial} = 0 \tag{2.37}$$

and therefore only $Y_{2,0}(\theta', \phi')$ remains. This simplifies equation 2.28 to:

$$\begin{aligned} D_{IS}^{axial}(\theta', \phi') = & d_{IS} \sqrt{\frac{16\pi}{5}} \left\langle Y_{2,0}(\theta', \phi') \right\rangle_{axial} \times \\ & \left[A_a d_{0,0}^{(2)}(\beta) + \sqrt{\frac{3}{8}} A_r \left(d_{0,-2}^{(2)}(\beta) e^{2i\gamma} + d_{0,2}^{(2)}(\beta) e^{-2i\gamma} \right) \right] \end{aligned} \tag{2.38}$$

By construction of a local frame with z'-axis parallel to the average vector orientation, we can identify α and β as [104]:

$$\beta = -\theta_{av} \qquad \text{and} \qquad \gamma = -\phi_{av} \tag{2.39}$$

and using Wigner matrices and spherical harmonics definitions (equations 2.24 and 2.27) equation 2.38 can be rewritten as:

$$\begin{aligned} D_{IS}^{axial}(\theta', \phi') = d_{IS} \sqrt{\frac{16\pi}{5}} & \left\langle \frac{3\cos^2\theta' - 1}{2} \right\rangle_{axial} \times \\ & \left[A_a \left(3\cos^2\theta_{av} - 1 \right) + \frac{3}{2} A_r \sin^2\theta_{av} \cos 2\phi_{av} \right] \end{aligned} \tag{2.40}$$

Introducing an order parameter S^2_{axial} as in equation 2.29, we obtain:

$$S^2_{axial} = \frac{4\pi}{5} \left\langle Y_{2,0}(\theta', \phi') \right\rangle^2_{axial} = \left\langle \frac{3\cos^2\theta' - 1}{2} \right\rangle^2_{axial} \tag{2.41}$$

and

$$D^{axial}_{IS} = S_{axial} D^{static}_{IS}(\theta_{av}, \phi_{av}) \tag{2.42}$$

This expression shows that the dynamic averaging effect of an axially symmetric motion depends only on the amplitude Θ, through the averaged value $\langle . \rangle_{axial}$ of the motion and not on the average orientation of the vector of interest in the PAS. For this reason this motion will be regarded as isotropic in the following work.

2.6.3 *Gaussian Axial Fluctuation Model*

BASIS OF THE MODEL. The Gaussian Axial Fluctuation (`GAF`) model was introduced by Bremi and Brüschweiler [41] to interpret protein backbone ^{15}N NMR relaxation data. Two different versions of this model exist; one with unidimensional reorientation, called `1D-GAF`, and one allowing for a three-dimensional reorientation process which is called `3D-GAF`. The assumptions of the model are the following:

- the peptide plane is considered as a planar object with a defined topology
- motion can be describe through diffusive reorientations around three orthogonal axes, defined using the average position of the peptide plane:
  1. the $\mathcal{A}_\gamma$ axis, defined by the two C^α atoms that flank the peptide plane
  2. the $\mathcal{A}_\beta$ axis, orthogonal to the peptide plane
  3. the $\mathcal{A}_\alpha$ axis, orthogonal to the two previous axes
- the three diffusive reorientations are assumed to be independent
- the diffusive reorientation follows a Gaussian distribution, centered on the average peptide plane position and whose standard deviation characterizes the amplitude of the motion. These amplitudes are named $\sigma_\alpha, \sigma_\beta$ and σ_γ for the three axes $\mathcal{A}_\alpha, \mathcal{A}_\beta$ and $\mathcal{A}_\gamma$ and describe the width of the Gaussian distributions at half height.

A peptide plane with associated GAF axes is shown in Figure 13. This representation of motion is probabilistic, in the sense that the motion is represented by the distribution of orientations and therefore no time-dependent trajectory can be extracted from the amplitudes alone.

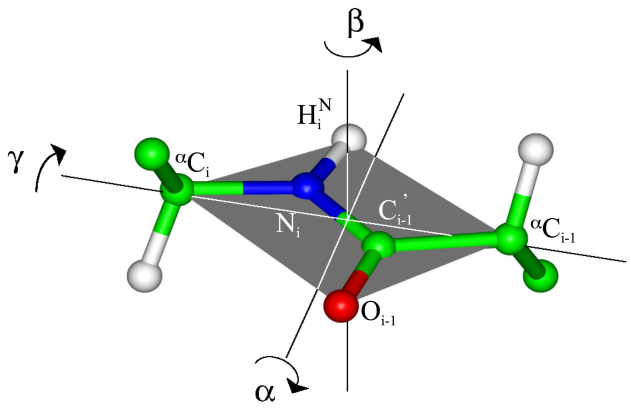

Figure 13 – Axes of peptide plane reorientation used in the GAF model. $\mathcal{A}_\alpha, \mathcal{A}_\beta$ and $\mathcal{A}_\gamma$ axes correspond to the indicated α-, β- and γ-motions.

The physical origin of the GAF motion can be found in the so-called Langevin oscillator, which corresponds to a system submitted to Brownian diffusion in a harmonic potential [105]. According to the Langevin formalism [106, 107] this can be described as a system submitted to three forces: the presence of the potential that restricts the diffusion process, a friction force and random thermal fluctuations. This gives rise to a stationnary solution, where the distribution of positions is a Gaussian centered at the equilibrium position [107].

1D-GAF AVERAGING. In this model the three reorientations are treated in a perfectly equivalent way and they will be therefore treated simultaneously in the following derivation by using a ι subscript which can represent α, β or γ.

The probability distribution $p_{\text{1D-GAF}}$ of finding a vector in a given orientation, for a unidimensional reorientation 1D-GAF, can be conveniently written in a local frame $\mathcal{R}_\iota$ whose z-axis is the axis of reorientation as:

$$p_{\text{1D-GAF},\iota}(\theta_\iota, \phi_\iota) = \frac{1}{\sqrt{2\pi\sigma_\iota^2}} \exp\left[-\frac{(\phi_\iota - \phi_{\iota,0})^2}{2\sigma_\iota^2}\right] \delta(\theta_\iota - \theta_{\iota,0}) \tag{2.43}$$

where $(\theta_\iota, \phi_\iota)$ and $(\theta_{\iota,0}, \phi_{\iota,0})$ describe the orientation and the mean orientation of the studied vector respectively, and where δ is the Dirac function.

As previously shown (see equation 2.23) an RDC can be expressed as a linear combination of second rank spherical harmonics, thus its average can be obtained by averaging the spherical harmonics.

As visible in equation 2.24, each of the five spherical harmonics can be expressed as:

$$Y_{2,p}(\theta,\phi) = P_{2,p}(\theta)\exp(ip\phi) \tag{2.44}$$

where $P_{2,p}(\theta)$ are the so-called associated Legendre polynomials of order 2 and degree p.

Using this decomposition it become possible to express the average spherical harmonics as:

$$\begin{aligned}\left\langle Y_{2,p}(\theta_\iota,\phi_\iota)\right\rangle_{\text{1D-GAF},\iota} &= \int_0^{\pi}\int_0^{2\pi} Y_{2,p}(\theta_\iota,\phi_\iota)p_{\text{1D-GAF},\iota}(\theta_\iota,\phi_\iota)\,d\phi_\iota d\theta_\iota \\ &= \int_0^{\pi} P_{2,p}(\theta_\iota)\delta(\theta_\iota-\theta_{\iota,0})\,d\theta_\iota \times \\ &\quad \int_0^{2\pi}\exp(ip\phi_\iota)\frac{1}{\sqrt{2\pi\sigma_\iota^2}}\exp\left[-\frac{(\phi_\iota-\phi_{\iota,0})^2}{2\sigma_\iota^2}\right]d\phi_\iota\end{aligned} \tag{2.45}$$

The second integral can be solved using [108]:

$$\int_{\mathbb{R}}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{x^2}{2\sigma^2}\right]\exp(iax)\,dx = \exp\left(\frac{-a^2\sigma^2}{2}\right) \tag{2.46}$$

where a and σ are real numbers. Here, the integration is only made for $\phi \in [0,2\pi]$, but if a motion of an amplitude σ_ι smaller than $\pi/3$ (60°) is assumed, due to the Gaussian modulation, the integration interval can be extend to $\mathbb{R}$ as the integrant will already have a negligible value in the upper and lower limits. This approximation physically correspond to neglecting the possibility for a peptide plane to make a complete reorientation about an axis. We thus obtain:

$$\left\langle Y_{2,p}(\theta_\iota,\phi_\iota)\right\rangle_{\text{1D-GAF},\iota} = P_{2,p}(\theta_{\iota,0})\exp(ip\phi_{\iota,0})\exp\left(\frac{-q^2\sigma_\iota^2}{2}\right) \tag{2.47}$$

which eventually gives:

$$\left\langle Y_{2,p}(\theta_\iota,\phi_\iota)\right\rangle_{\text{1D-GAF},\iota} = Y_{2,p}(\theta_{\iota,0},\phi_{\iota,0})\exp\left(\frac{-q^2\sigma_\iota^2}{2}\right) \tag{2.48}$$

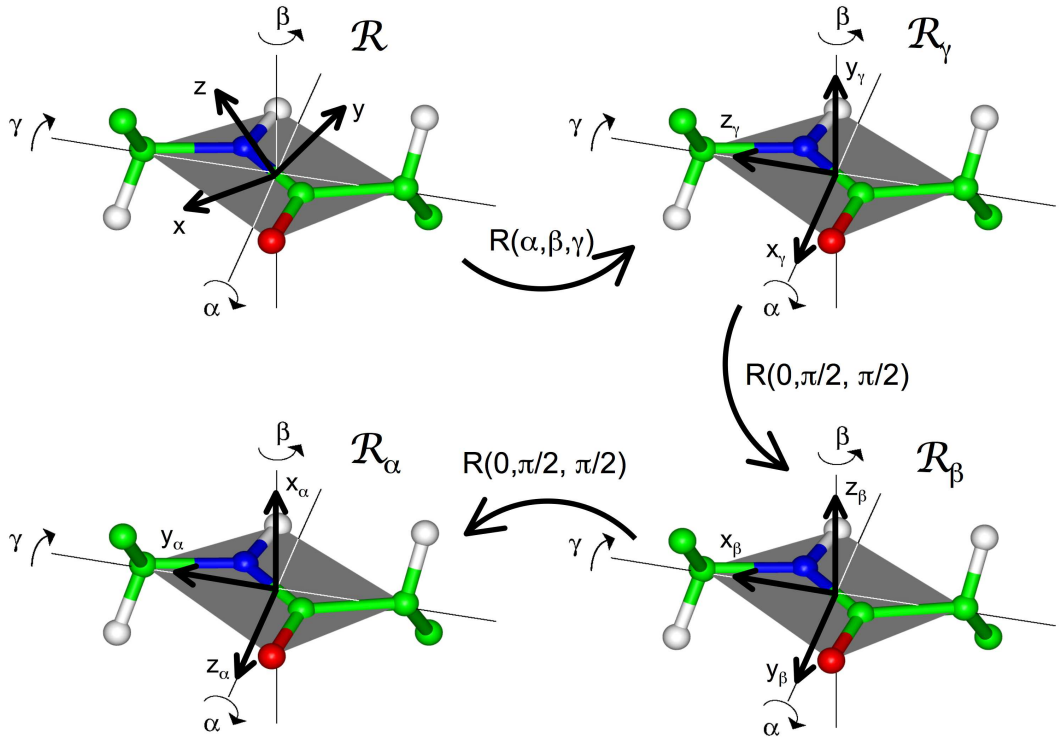

Figure 14 – Euler rotations used during 3D-GAF dynamical averaging. The first rotation corresponds to switching from the PAS to the $\mathcal{R}_\gamma$ frame. This rotation is used for γ-1D-GAF or 3D-GAF motion (a similar rotation allows the characterization of other 1D-GAF motions). Two supplementary rotations are necessary to achieve 3D-GAF averaging, by switching to the $\mathcal{R}_\beta$ and $\mathcal{R}_\alpha$ frames.

This equation has been expressed in the local $\mathcal{R}_\iota$ frame and can be expressed in an alternative frame $\mathcal{R}$ using equations 2.25:

$$\left\langle Y_{2,p}(\theta,\phi)\right\rangle_{\text{1D-GAF},\iota} = \sum_{q=-2}^{2} D^{(2)}_{q,p}(\alpha_\iota,\beta_\iota,\gamma_\iota)\exp\left(\frac{-q^2\sigma_\iota^2}{2}\right) Y_{2,q}(\theta_\iota,\phi_\iota) \quad (2.49)$$

If we consider that $\mathcal{R}$ is the PAS, and $(\alpha_\iota,\beta_\iota,\gamma_\iota)$ are the three Euler angles that describe the rotation between $\mathcal{R}$ and the local frame $\mathcal{R}_\iota$, this expression corresponds to the averaging of the spherical harmonics in the PAS, with the motion described in the local frame $\mathcal{R}_\iota$ (see Figure 14). This can be directly used for estimating the motional averaging through 1D-GAF motion as:

$$D_{IS}^{\text{1D-GAF},\iota} = d_{IS}\sqrt{\frac{16\pi}{5}}\Bigg[A_a\left\langle Y_{2,0}\right\rangle_{\text{1D-GAF},\iota} + \sqrt{\frac{3}{8}}A_r\left(\left\langle Y_{2,-2}\right\rangle_{\text{1D-GAF},\iota} + \left\langle Y_{2,2}\right\rangle_{\text{1D-GAF},\iota}\right)\Bigg] \quad (2.50)$$

with

$$\left\langle Y_{2,p}(\theta,\phi)\right\rangle_{\texttt{1D-GAF},\iota} = \sum_{q=-2}^{2} e^{ip\alpha_\iota} d^{(2)}_{q,p}(\beta_\iota) e^{ip\gamma_\iota} \exp\left(\frac{-q^2\sigma_\iota^2}{2}\right) Y_{2,q}(\theta_{\iota,0},\phi_{\iota,0}) \tag{2.51}$$

3D-GAF AVERAGING. The 3D-GAF motional averaging can easily be deduced from the previous calculation. As the three reorientations are independent their averaging effects can be successively applied, for a given spherical harmonic, we obtain:

$$\left\langle Y_{2,p}(\theta,\phi)\right\rangle_{\texttt{3D-GAF}} = \left\langle\left\langle\left\langle Y_{2,p}(\theta,\phi)\right\rangle_{\texttt{1D-GAF},\alpha}\right\rangle_{\texttt{1D-GAF},\beta}\right\rangle_{\texttt{1D-GAF},\gamma} \tag{2.52}$$

This can be developed by iteratively applying equation 2.51. The Euler angles for the two rotations from $\mathcal{R}_\gamma$ to $\mathcal{R}_\beta$ and $\mathcal{R}_\beta$ to $\mathcal{R}_\alpha$ are $(0,\pi/2,\pi/2)$ as switching from one frame to another corresponds to an appropriate permutation of the three axes. The three successive frame changes are illustrated in Figure 14.

$$\begin{aligned}\left\langle Y_{2,p}(\theta,\phi)\right\rangle_{\texttt{3D-GAF}} = \sum_{q=-2}^{2}\Bigg[& e^{-iq\alpha}\, d^{(2)}_{q,p}(\beta)\, e^{-ip\gamma}\, e^{-\frac{1}{2}q^2\sigma_\gamma^2} \\ & \times \sum_{r=-2}^{2}\Bigg[d^{(2)}_{r,q}\left(\frac{\pi}{2}\right) e^{-iq\frac{\pi}{2}}\, e^{-\frac{1}{2}r^2\sigma_\beta^2} \\ & \times \sum_{s=-2}^{2}\Bigg[d^{(2)}_{s,r}\left(\frac{\pi}{2}\right) e^{-ir\frac{\pi}{2}}\, e^{-\frac{1}{2}s^2\sigma_\alpha^2} \\ & \times Y_{2,s}(\theta_{\alpha,0},\phi_{\alpha,0})\Bigg]\Bigg]\Bigg]\end{aligned} \tag{2.53}$$

and the corresponding RDC is obtained through:

$$D_{IS}^{\texttt{3D-GAF}} = d_{IS}\sqrt{\frac{16\pi}{5}}\left[A_a\left\langle Y_{2,0}\right\rangle_{\texttt{3D-GAF}} + \sqrt{\frac{3}{8}}A_r\left(\left\langle Y_{2,-2}\right\rangle_{\texttt{3D-GAF}} + \left\langle Y_{2,2}\right\rangle_{\texttt{3D-GAF}}\right)\right] \tag{2.54}$$

2.7 CONCLUSION

This chapter aimed at presenting some experimental aspects, the origin and the relevance of RDCs for studying biomolecular systems. Under favorable

conditions, it is often possible to measure different sets of RDCs for a given system. The information content of these datasets is significant, in the sense that they simultaneously provide an accurate source of both structural and dynamic information.

The structural information is mainly an orientational information that can be highly degenerate but that measures the relative orientation of each probed vector towards a common alignment tensor.

In this chapter, RDCs were also presented as a source of motional information. This information is highly complementary to that extracted from spin relaxation data (see Chapter 1).

Both give rise to tensorial phenomena, the relaxation through the tensor of diffusion that can be derived from the rotational diffusion properties of the system and the RDCs through the alignment tensor. One major difference between RDCs and spin relaxation rates is that relaxation forces the system to evolve through random fluctuation of the spatial organization of the system, inducing random transitions between spin states at rates that are dependent on the frequency of the motion. This makes relaxation data sensitive to the different frequencies induced by the motion. This characteristic allows to extract characteristic timescales from relaxation measurements which is not possible with RDCs. Nevertheless this also prevents the extraction of information about motional timescales slower than the correlation time of the molecule as the overall rotational motion will already have scaled down the correlation function to zero.

Comparing those two sources of information can be achieved by comparing order parameters. For relaxation data this order parameter is the plateau value of the internal correlation function, whereas for RDCs which are "time-insensitive" it correspond to the degree of order that remains when considering the existence of internal dynamics. RDC order parameters are sensitive to longer timescales and thus comparing relaxation and RDC order parameters allows the characterization of both fast (faster than the correlation time) and slow (slower than the correlation time) dynamics.

Part III

FOLDED PROTEIN DYNAMICS

3

OVERVIEW OF DYNAMIC DESCRIPTIONS OF RDCS

ABSTRACT

RDCs are powerful probes of protein dynamics occurring on timescales up to the millisecond. Various techniques have been developed in recent years in order to extract this information and this chapter aims at reviewing these approaches in the scope of the work presented later in the Thesis. Emphasis is placed on descriptions that allow the characterization of site specific dynamics using numerical or analytical models such as ensemble averaging approaches, model-free descriptions and GAF modeling. The use of the GAF description of the peptide plane reorientation is extensively presented as it constitutes the basis of much of the analytical work presented in this Thesis.

3.1 INTRODUCTION

As discussed in Chapter 2, RDCs are excellent probes of biomolecular motions. They provide site specific information and their measurement in different alignment media leads to different perspectives of the same motion taken from different points of view.

It is worth emphasizing that the dynamics discussed here corresponds to motions with characteristic timescales faster than the millisecond. They are therefore not observable through real time NMR, even if recent developments reduced the lower limit of timescales accessible through fast NMR approaches to timescales of the order of the second [109, 110]. Motions occurring on real-time NMR accessible timescales are often straightforward to analyze and can be seen as a time-dependent sequence, like a movie, in the sense that the experimental dataset corresponds to successive states of the system.

With RDCs, the dynamics can be interpreted only via measurements that average over the entire system dynamics. It is therefore impossible to develop

a time-dependent trajectory based description of the system uniquely from the experimental data. In the previous chapter we briefly presented some tools that can be used to extract dynamical parameters from experimental RDCs. This chapter will first briefly discuss some prerequisites to a dynamic analysis of RDCs, before turning to the description of different kinds of approaches, based on domain tensor comparison, multi-structures ensemble descriptions and analytical motional models. The emphasis will be put on folded protein dynamics, even if some applications in other biomolecular systems will be mentioned.

The forthcoming discussion concerns results published by different authors. For the consistency of notation some liberty has been taken to modify existing notations presented in the original publications.

3.2 FROM STRUCTURE TO DYNAMICS

Historically the majority of publications treating RDCs measured in biomolecules assume a so-called static description that supposes the absence of differential dynamics within the framework of the molecular or alignment frame. This is obviously an approximation in the sense that a general physical description such as the energy equipartition principle [107], implies that thermal energy is distributed in the large number of different motional modes available, inducing interconversions between the numerous substates that one would expect to exist in a complex macromolecule such as a protein. However following a number of initial studies on model proteins for which a high resolution structure was already known, it was clear that a very good reproduction of experimental data was often obtained by invoking only a static structure. This suggested that the amount of dynamics present was either minimal, or averaged in such a way as to remain in agreement with a static description. Nevertheless, as understood by one of the very first applications of RDCs to the study of proteins [111], important dynamic information can be extracted from these quantities.

3.2.1 *Structure or Dynamics?*

One of the most detailed applications of RDCs to the understanding of molecular structure was presented by Bax and co-workers in 2003, when they published the analysis of a large RDC dataset (multiple couplings from each peptide plane) measured in five different alignment media in protein GB3 [112]. These data were used to analyze the structure and dynamics of individual peptide planes. The planarity of the peptide plane has been shown to be on average a reasonable approximation, but high resolution

X-ray crystallography often shows significant deviations from this ideal geometry [113]. In this case several RDCs were measured within each peptide plane although agreement with a refined structure was good using optimal geometry for the peptide plane, the authors nevertheless tried, in a first step, to reduce this discrepancy by optimizing local geometry.

The study showed that real improvement in data reproduction could be achieved using only structural refinement and that the N_i-H_i^N bond vector differed slightly from the idealized geometry where the N_i-H_i^N vector is assumed to be in plane and along the line bisecting the C_i^α-N_i-C'_{i-1} angle.

Nevertheless, RDCs for some residues could not be explained using only structural optimization which in terms of the degenerate orientations discussed in the previous chapter, corresponds to the observation that the different orientational solutions for the measured RDCs could not intersect within the experimental error. The presence of dynamics was therefore investigated and for N_i-H_i^N vectors it was suggested that dynamics may occur out of the peptide plane.

The study underlined the fact that distinction between structural and dynamic features is not always obvious when analyzing RDCs and that dynamic descriptions, which require more parameters to define their physical properties, need to be investigated statistically in order to test for significance.

The insufficiency of a single structural model to adequately describe experimental data from multiple alignment media may also stem from the changes in the molecular environment associated with the different liquid crystal media. It is therefore important, for both structural and dynamic analyses, to be able to check that the data are not being affected by the alignment medium and thereby biasing the interpretation of conformational behaviour. Below an approach that addresses this consistency will be presented.

3.2.2 SECONDA *Analysis*

The Self-Consistency of Dipolar Couplings Analysis (SECONDA) is an approach that can estimate the level of self-consistency of a set of RDCs measured in multiple alignment media [114, 115]. This is based on the analysis of the weighted covariance matrix $\mathbf{C}$, defined as:

$$C_{ij} = \frac{1}{M-1} \sum_{k=1}^{M} \frac{1}{\sigma_k^2} \left(D_i^{(k)} - \left\langle D_i \right\rangle \right) \left(D_j^{(k)} - \left\langle D_j \right\rangle \right) \tag{3.1}$$

where i and j run over the N different RDCs experimentally measured in all of the M alignment media and where:

$$\left\langle D_i \right\rangle = \frac{1}{M} \sum_{k=1}^{M} D_i^{(k)} \tag{3.2}$$

$$\sigma_k = \frac{1}{N-1} \sum_{i=1}^{N} \left(D_i^{(k)} - \left\langle D_i \right\rangle \right)^2 \tag{3.3}$$

which corresponds respectively to the average of the i-th coupling in all the alignment media and the variance of the RDCs measured in the k-th alignment medium.

A principal component analysis of this matrix would, in the absence of noise, reveal only five non zero eigenvalues if all RDCs derive from a single structure or an ensemble of structures, that are identical in all media [114]. This is equivalent to saying that the data can be expressed as a combination of RDCs measured in five linearly independent media. Therefore, the appearance of more than five non zero eigenvalues can be only be due to experimental noise or inconsistency in the experimental datasets. Equivalent information can be obtained through a SVD decomposition of a matrix constructed with all RDCs measured in the different alignment media [116] (see Annexe A).

This approach can be used to check the consistency of a dataset or to select within a large dataset a sub ensemble with higher self-consistency [115] and can be used as an initial check in order to avoid interpreting inconsistencies as physical properties.

3.3 FRAGMENT TENSOR ANALYSIS

It is possible to characterize a single rigid structure using a single alignment tensor (see Section 2.5). However, the analysis of the alignment tensor can be independently done for different fragments of the system, as long as enough RDCs (5 in theory) are available for each fragment and the structure of the fragment is sufficiently well known. If the system is a rigid object and if the assumed relative orientations of the fragments are correct all the alignment tensors should converge within the experimental accuracy.

Nevertheless if one of those two assumptions are incorrect, discrepancies between the alignment tensors can be found. For example variations in the relative orientations of the tensors — that is to say (α, β, γ) Euler angles — can be rationalized by modifying the relative orientation of one domain

compared to the others [19, 48]. Changes in magnitude and rhombicity of the tensors can then be interpreted in terms of dynamics. This is the basis of fragment tensor analysis where studied fragments can be for example a complete domain of a biomolecule or a smaller fragment.

3.3.1 *Domain Motions*

The now established importance of RDCs as motional probes was originally revealed during the analysis of magnetically aligned cyanometmyoglobin [111]. In this study, RDCs were measured at three different fields (and therefore three degrees of alignment as it was magnetically induced) and the accuracy of the measurements was good enough to ensure that the discrepancy between measured and estimated data (using a X-ray structure) was systematic and meaningful. This discrepancy was reduced by invoking the presence of motions for α-helices of fixed geometry, through a model of diffusion in a cone or diffusion around a single axis.

Following this pioneering study, a large number of systems have been shown to exhibit domain-like dynamic behavior using similar approaches, for example, the analysis of Barley Lectin domain [117] required the use of a diffusion in a cone model with a semi-angle of $\sim 40^\circ$ in order to explain the large variations in the A_a and A_r components of the two studied domains. Other studies have followed concerning various kinds of systems such as proteins [118–120], RNA [121, 122] or oligosaccarides [123].

It is not always obvious to attribute such discrepancies to dynamic averaging [124]. Under certain conditions the dynamic modulation of the alignment tensor can be invisible due to a correlation between alignment tensor changes and local dynamics [125] and in some cases complementary sources of information have to be used in order to confirm the validity of the dynamic description [126–128].

3.3.2 *Local Alignment Tensor*

The idea of this analysis is to push the fragment analysis to its natural limit by using the smallest fragment possible [129]. Nevertheless this approach requires at least five RDCs per fragment to determine an alignment tensor and more are required to ensure the robustness of the approach. As previously shown the peptide plane is the natural rigid fragment that can be found in the protein backbone but due to its planarity and the limited number of measurable couplings, larger fragments have to be used. Tolman et al. proposed to use the peptide plane and the subsequent tetrahedral junction as structural motif in order to attain up to eight RDCs per unit.

Clearly as the orientation of the tetrahedral junction has to be determined site specifically (it depends on the backbone ϕ-angle) a structural input, such as X-ray or NMR structure, or experimentally determined scalar couplings, is needed to fix each fragment topology. Here, a generalized degree of order (GDO) which can be seen as a norm of the alignment tensor, was introduced in order to characterize the fragment dynamics, as:

$$\vartheta = \sqrt{\frac{2}{3}\sum_{ij} A_{ij}^2} \tag{3.4}$$

The GDO, which can be a convenient measure of dynamics in the case of isotropic motions can be more difficult to interpret for anisotropic motions, in the sense that it will depend on both the amplitude of the motion and the fragment orientation. Nevertheless the approach was successfully applied to different systems, including sugar moieties [123].

3.4 ENSEMBLE AVERAGED APPROACHES

Ensemble averaging can be applied in various applications, as all approaches that describe dynamics using an ensemble of conformers make use of this description. Nevertheless, different applications can be very different in terms of philosophy. Three different classes of methods can be distinguished:

- those using on an ensemble average where the conformer ensemble is actively refined against experimental data and force fields
- those based purely on molecular dynamics: using an ergodic hypothesis an ensemble description can be obtained from a molecular dynamics trajectory
- those where large conformational sampling is defined before selecting a sub-ensemble on the basis of the experimental data

One of the major interests in those approaches is the breath of their application range: they can be used to describe systems where different kinds of experimental data are available (RDCs, PREs, nOes...) and applied to a very broad range of systems from folded to unfolded proteins.

3.4.1 *Ensemble Restrained Molecular Dynamics*

This approach was pioneered for the use of NMR in the late 1980's [130–134] and first applied to RDCs by Clore and Schwieters. Ensemble averaged

restrained molecular dynamics is a direct extension of classical NMR structure refinement. The philosophy of this approach is to consider that instead of refining a single structure, it might be more relevant to refine simultaneously different structures and apply experimental restraints, not to each of the conformers but to the complete ensemble. Therefore experimental restraints have to be in agreement on average, which give less restrictive definition of the accessible conformational landscape. Nevertheless each conformer cannot be considered meaningful by itself, and thus the only way to correctly interpret the results is through a statistical interpretation of the obtained ensemble. The use of few structures can be problematic for such a statistical analysis and therefore calculations are sometimes repeated and averaged over a large number of calculations.

Initial applications include protein GB3 [135] and Ubiquitin [136] using RDCs and nOe experimental restraints. In both cases, the number of statistically relevant conformers were at most two implying a two-site jump where differences in conformations were required. The statistical significance of introducing this second conformer was not always easy to demonstrate. Further applications were proposed for nucleic acids [137].

Concerning Ubiquitin, other ensembles were proposed, for example the one where ^{15}N relaxation order parameters and nOes were used as restraints [138] or the EROS ensemble of 116 structures, determined using an original ensemble average of 8 structures on the basis of a large set of RDCs and nOes [139].

This kind of approach has been applied to urea-denatured Ubiquitin, where eight conformers were found to be enough to reproduce PRE and RDC data [140]. This approach was also applied to RNA (ribonucleic acid) and the obtained conformers were used to a further interpretation in terms of motion trajectories [141].

3.4.2 *Molecular Dynamics Approach*

Molecular dynamics (MD) simulations are by nature a methodology adapted to studying molecular motions, relying on the prediction of the time evolution of a molecular system via Newton's dynamics [142, 143].

Therefore, if force fields are well parametrized [144] and the sampling large enough to correspond to the timescale of the studied data [142], a MD trajectory should in principle reproduce both the structural and dynamic feature of the system of interest.

Concerning experimental data such as ^{15}N relaxation which is sensitive to timescales shorter than the correlation time of the molecule (~5 -20 ns) the sampling issue can be easily overcome using currently available computational power. As a testimony to the fact that state of the art force fields have made considerable progress in recent decades, relaxation data can often be accurately reproduced using computationally accessible trajectories.

For RDCs, the sampling issue is much more problematic, as RDCs are sensitive to motions occurring on timescales up to the millisecond. In order to properly sample such a time window, a trajectory of tens of milliseconds would have to be used, which is not currently computationally possible [142]. This sampling issue gives rise to a further force field parameterization issue in the sense that pushing sampling limits further than the currently accessible hundreds of nanosecond to microseconds range has to be achieved using force fields that would need to be optimized to reproduce longer timescales.

Currently, two options are available to study RDC dynamics with MD. The first one is to make MD as long as possible and this will give a satisfying description of the system if negligible motions occur on timescales longer than the MD trajectory. The second is to enhance artificially the sampling by using one of the numerous ways that have been proposed to accelerate the events occurring in molecular dynamics trajectories [145–147].

An ensemble of structures resulting from a long MD simulation can be used to directly interpret experimental data for example by estimating RDC reproduction using a SVD based analysis [148] (see Annexe A) or can be analyzed analytically to extract parameters that can be compared to analytical approaches such as the `3D-GAF` model [149].

3.4.3 *Sample and Select*

The sample and select approach was proposed by Chen et al. [150] and is based scheme is:

1. the conformational landscape accessible to the system is extensively sampled in order to obtain a large ensemble of conformers that is supposed to reproduce all the imaginable conformations of the system.

2. from this conformer database a sub-ensemble is selected to reproduce experimental data.

The definition of the database is a crucial issue for this approach as it will define the available conformational space of the system. In other words any

relevant conformation missing in this starting database will be definitively lost for the analysis.

This database can be built using any approach that allows a broad sampling of conformational space, such as molecular dynamics or Monte-Carlo approaches.

In the second step structures are selected from the database in order to reproduce experimental data, but no optimization or refinement of the selected structures is applied. This is a major difference compared to restrained ensemble averaged molecular dynamics, which is also responsible for the discrepancy in the number of conformers necessary to reproduce data. Sample and select methods normally require more structures but the relevance of each structure is ensured by the absence of optimization against experimental data. It is worth emphasizing that, as with other ensemble averaged method, this approach does not confer physical meaning to a single conformer and thus needs to be analyzed within a statistical description. Nevertheless, the need for larger ensembles provides a more intuitive basis for statistical interpretations.

This approach has been applied to a large set of systems, including folded proteins [150], unfolded proteins [151] or nucleic acids [152].

3.5 MODEL-FREE ANALYSIS

As for relaxation data analysis, the idea of a model free analysis is to describe motional averaging without invoking any kind of motional model. This avoids bias due to an inevitably imperfect model, but limits the possible interpretation of the results as most of the obtained parameters are purely mathematical and therefore sometimes difficult to interpret physically.

Two different model free approaches have been developed in order to extract dynamic information content from RDCs. Both are based on matrix descriptions of RDCs measured in different alignment media and lead to similar information about structure and dynamics, however the used formalism differ significantly.

3.5.1 *Self-Consistant RDC-based Model-Free Approach*

As shown in equation 2.28, it is possible to express a given RDC in an arbitrary molecular frame, as a function of dynamically averaged spherical harmonics. If RDCs are measured in different alignment media, it becomes possible to rewrite this equation as a system of linear equations. Knowledge

of tensor properties and averages of the spherical harmonics is enough to fully characterize the structural and dynamical properties of the system (orientational envelope of the sampling for the vector of interest). This is the basis of this model-free approach [104].

As applied, this approach has been shown to determine the relative amplitude and nature of dynamics throughout the protein, without estimating the absolute level of alignment (and therefore dynamics) [104], that is to say the main component of the alignment tensor A_{zz}. Changing the value of the principal component will scale the determined spherical harmonics average values and an overall scaling factor $S_{overall}$ has to be applied during the analysis to obtain meaningful parameters (for example by comparing to order parameters determined from spin relaxation). This is equivalent to considering the ${}^1D_{NH}$ couplings in arbitrary units, corresponding to $A_{zz} = 1$ and noted $\overline{D}_{i,m}$, where i is the residue number and m run over the different alignment media.

Thus, the method mainly consists of solving the following system of equations:

$$\overline{D}_{i,m} = \sum_{l=-2}^{2} F_l(R_m, \alpha_m, \beta_m, \gamma_m) \left\langle Y_{l,q}(\theta'_i, \phi'_i) \right\rangle_I \tag{3.5}$$

where $F_l(R_m, \alpha_m, \beta_m, \gamma_m)$ is a coefficient depending on the rhombicity R_m, the orientations $(\alpha_m, \beta_m, \gamma_m)$ of the considered tensor and (θ'_i, ϕ'_i) the orientation of the N_i-H_i^N vector in a common molecular frame.

In order to obtain information about tensor properties, which will define the values of the F_l coefficients a structural model is used. The dependence on this first structural input can be reduced by using an iterative process where the resolution of the previous step is used at each subsequent step and where the obtained averaged orientation is used to establish a new tensor [153].

By construction of this system of equations, five independent equations and therefore five independent alignment media are needed in order to obtain a unique solution.

In order to facilitate the interpretation of the results, the averaged spherical harmonics are re-expressed in a frame $\mathcal{R}''$ where $\left\langle Y_{2,0}(\theta''_i, \phi''_i) \right\rangle_I$ is maximal, which corresponds to the local frame in which z''-axis is collinear to the center of the distribution of orientations of the studied N_i-H_i^N vector. The orientation of the axis provides structural information as it corresponds to the mean orientation $(\theta'_{eff}, \phi'_{eff})$ of the studied vector [104]. Order parameters can be estimated using 2.29 and the anisotropy of the motion can be

obtained from two parameters, η that characterizes the amplitude of the anisotropy and ϵ that defines the orientation of this anisotropy:

$$\eta = \sqrt{\frac{\sum_{p=\{-2,2\}} \left\langle Y_{2,p}(\theta_i'', \phi_i'') \right\rangle_I \left\langle Y_{2,-p}(\theta_i'', \phi_i'') \right\rangle_I}{\sum_{p=\{-2,-1,0,1,2\}} \left\langle Y_{2,p}(\theta_i'', \phi_i'') \right\rangle_I \left\langle Y_{2,-p}(\theta_i'', \phi_i'') \right\rangle_I}} \tag{3.6}$$

$$\epsilon = \frac{1}{2} \arctan \frac{\left\langle Y_{2,+2}(\theta_i'', \phi_i'') \right\rangle_I - \left\langle Y_{2,-2}(\theta_i'', \phi_i'') \right\rangle_I}{i\left(\left\langle Y_{2,+2}(\theta_i'', \phi_i'') \right\rangle_I + \left\langle Y_{2,-2}(\theta_i'', \phi_i'') \right\rangle_I \right)} \tag{3.7}$$

These approaches were initially developed by Griesinger, Brüshweiler and co-workers using simulated data form molecular dynamics simulations [104] and then applied to experimental data from the protein Ubiquitin with an increasing number of RDC datasets [154–157]. The latest version of this analysis called SCRM for Self-Consistent RDC-based Model-Free approach [153], included a SECONDA analysis [115] in order to select a highly coherent sub-ensemble of 23 of the over 36 different RDC datasets.

3.5.2 *Direct Interpretation of Dipolar Couplings*

This approach developed by Tolman [158] is based on the description of local dynamics with a local matrix order, similar to the Saupe matrix for the alignment.

In order to simplify equation 2.15 an order matrix, the Saupe matrix, was introduced to characterize global reorientation of the $\mathbf{B}_0$ field in a molecular frame. Nevertheless, this expression is perfectly symmetric for global and local motion and therefore the reorientation of the studied internuclear vector in this molecular frame can be studied in an entirely equivalent way. Similarly to equation 2.16, a local order matrix $\hat{\mathbf{B}}$ is introduced as:

$$B_{kl} = \frac{1}{2}\left(3 \left\langle \cos \alpha_k \cos \alpha_l \right\rangle_I - \delta_{kl}\right) \qquad k, l \in \{x, y, z\} \tag{3.8}$$

If now the five independent elements of the Saupe matrix for the alignment medium m $\hat{\mathbf{A}}_m$ are placed in a column vector $\mathbf{A}_m$ and the local order matrix for the vector i $\hat{\mathbf{B}}_i$ as a line vector $\mathbf{B}_i$, the expected value of the RDC i in the medium m is given by:

$$D_{i,m} = 2\, d_{IS}\, \mathbf{B}_i \mathbf{A}_m \tag{3.9}$$

The generalization for N RDCs measured in M alignment media is straightforward, by introducing $\mathbf{B}$ a N×5 matrix whose i-th line is $\mathbf{B}_i$ and $\mathbf{A}$ a 5×N

matrix whose n-th column is $\mathbf{A_m}$ and $\mathbf{D}$ a $N \times M$ matrix containing all RDCs. All RDCs can be estimated through the following matrix product:

$$\mathbf{D} = 2\, d_{IS}\, \mathbf{BA} \tag{3.10}$$

In this approach, the author attempts to circumvent the determination of $\mathbf{A}$. This can be achieved by using at least five independent alignment media, which means that $\mathbf{A}$ is a rank 5 matrix. In this case $\mathbf{D}$ and $\mathbf{B}$ matrices have the same rank, which here means that their image can be generated using two orthonormal bases related by a unitary transform $\mathbf{T}$. Thus, the singular value decomposition of $\mathbf{B}$ can be written as (see Annexe A):

$$\mathbf{B} = \mathbf{U_B W_B V_B^T} = \mathbf{U_D T W_B^T V_B} = \mathbf{U_D \Lambda} \tag{3.11}$$

$\mathbf{U_B}$ and $\mathbf{U_D}$ are built with the range-spanning basis of $\mathbf{B}$ and $\mathbf{D}$. The 5×5 $\mathbf{\Lambda}$ matrix describes how to produce the image of $\mathbf{B}$ from the range-spanning basis of $\mathbf{D}$. This matrix cannot be deduced only from RDCs ($\mathbf{D}$) alone. To face this underdetermination the author proposed to use the solution that minimizes the variation in generalized order parameters as the solution has to describe the local structural and dynamic properties of the structured molecule.

This allowed the characterization of the $\mathbf{B}$ matrix and therefore the local structural and dynamic information. This information can be seen as equivalent to the one from the SCRM approach, as it becomes possible to extract the direction of the mean vector, amplitude of the motion and the direction and magnitude of the asymmetric component of the motion.

This method was again applied to the protein Ubiquitin using two different RDC datasets comprising 9 and 11 different alignment media. In the first case a structural input was used to stabilize the analysis [158] , and in the latter analysis [159] no structural input was used. This resulted in the accurate determination of the structural properties and the determination of order parameters, which appear to be more robustly determined than the anisotropic properties of the motion, and that again needs to be scaled using a uniform factor, in this case in order to be lower than order parameters extracted from ^{15}N relaxation. This approach was applied in a very similar form to the study of protein GB3 [160].

3.6 GEOMETRIC DESCRIPTION OF MOTION USING THE GAF MODEL

In addition to the described model-free approaches, the use of a geometric model based on biophysical principles can be a useful, and a complementary, way to characterize protein backbone dynamics. No biophysical model can

perfectly describe protein backbone motion, which is surely a complex and multi-faceted phenomenon, but use of a general, and hopefully appropriate model can lead to a relevant and informative interpretation of the dynamic averaging experienced by RDCs.

There are various advantages and drawbacks of using a specific biophysical model:

- Model-free approaches require at least five linearly independent alignment media in order to extract dynamic information. The use of a model with a low number of adjustable parameters can *a priori* circumvent this issue. However, the model must be flexible enough to allow an accurate characterization of possibly complex motions.

- Parameters extracted from such a model have a well-defined physical meaning and can be more easily interpreted than those derived from model-free approaches. Equally the use of an inappropriate model can lead to a misinterpretation of the data that will be less severe with model free approaches.

- The use of a local structural motif whose conformation is known allows the measurement of several RDCs in a given alignment medium. This combination of the dynamic information of multiple vectors leads to an increase of the possibility of accurately defining the dynamics. Conversely the incorporation of a structural motif in the model has to be consistent with the dynamic model that the motif can be expected to experience.

All the work presented in this section is based on the GAF model (see Section 2.6). The model was originally developed for the interpretation of spin relaxation data, on the basis of molecular dynamics simulations. This combination was used to demonstrate the capacity of the model to describe backbone motion in peptides and model proteins. Its development for the interpretation of RDC data was performed by Blackledge and co-workers, and the following results represent a large part of the basis on which the research work developed in this manuscript is based.

3.6.1 *Pioneering Work*

Following spin relaxation studies, where the unidimensional version of the GAF model was shown to capture most of the fast motional dynamics present in the protein backbone, a model called ortho-GAF was developed where the single axis of reorientation was set to be orthogonal to the NH

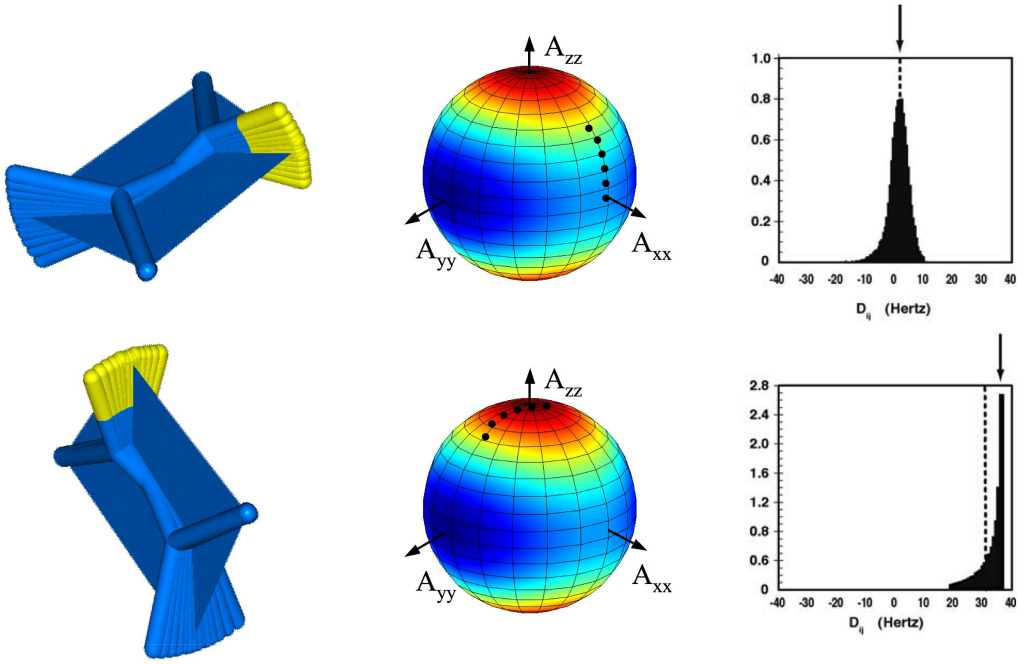

Figure 15 – Effect of GAF motion for dynamic averaging. Here two planes with the same reorientation amplitude but different orientations are presented (left part). The orientations sampled during the GAF motion are depicted with dots on the colored sphere (central part). Corresponding distributions of the RDCs $^{1}D_{NH}$ are shown (right part): the dotted line indicates the static RDCs corresponding to the mean position of the N_i-H_i^N internuclear vector, the arrow indicates the motionally averaged RDC.

internuclear vector [161]. Using simulated data the authors showed that even using a common amplitude for all the peptide planes, the motion characterized by its mean orientation and its direction will exhibit very different averaging properties due to this anisotropic behavior. This is in stark contrast to isotropic local motion — diffusion in an isotropic cone — where all RDCs would be scaled by the same factor, and therefore would reproduce the experimental data identically irrespective of the direction of this motion. For example rare cases were identified whereby the motional averaging can lead to a higher value for the dynamically averaged value of an RDC than the value predicted in the absence of motion. This situation is of course impossible for an isotropic motion such as an isotropic diffusion within a cone. This anisotropic averaging behavior is illustrated in Figure 15.

Applied to experimental data, for diverse proteins, it was shown that a common amplitude of GAF motion in secondary structural elements of $\sigma_o \sim 15°$ led to a statistically significant improvement of the data reproduction. Analysis was made using a single RDC dataset combined with a high resolution protein structure. This was the first demonstration that anisotropic motions

could be detected with statistical certainty and provided an important indication that this model was relevant to the averaging properties of backbone RDCs in proteins. In addition, comparing N_i-H_i^N and C_{i-1}^{α}-C'_{i-1} averaging behavior, a rationalization of the previously proposed use of a longer N_i-H_i^N bond than generally accepted (1.04 Å instead of 1.02 Å) was identified. The longer distance had indeed been proposed in order to account for the `1D-GAF` motion using a static approach. Here, the anisotropy of the motion suggested a value around 1.02 Å closer to the one commonly used for other studies such as ^{15}N relaxation.

This approach was then used to study local dynamics of protein GB3 and lysozyme, where at least two different alignment media RDC dataset were available [162].

The approach was then extensively tested in order to determine the sensitivity of this model against structural noise, and an important robustness of the `GAF` model compared to other descriptions such as the axially symmetric description [163] was identified. This study further underlined the importance of an accurate alignment tensor determination.

3.6.2 *3D-GAF Analysis*

The first complete `3D-GAF` analysis was achieved during the analysis of RDCs from protein GB3 [164]. In this case the analysis was made using a high resolution X-ray structure and the tensors were determined using the less dynamic RDCs. This study allowed the determination of the three reorientation amplitudes σ_α, σ_β and σ_γ.

The results revealed an alternating pattern for the amplitude of the γ- and β-motions in the β-sheet, where less dynamic residues were the hydrophobic ones. The amplitudes of motion increased from one edge of the sheet to the other, leading to maximal dynamic sampling in the interaction site. Correlation of the parameterized motions across the β-sheet were revealed using $^3J_{C'N}$ trans-hydrogen-bond scalar couplings. The analysis was supported by extensive cross validation.

3.6.3 *Simultaneous Determination of Structure and Dynamics*

This analysis was carried out on a similar dataset to the first `3D-GAF` analysis, but with an optimized tensor determination protocol. This analysis did not make use of a previously known structure, thereby removing structural bias. The only structural input was the idealized peptide plane topology, obtained as an average of the peptide planes of ultra-high resolution X-

ray and neutron structure[1] for the heavy atoms. Amide protons were placed in-plane according to an idealized direction determined from neutron structures — where deutrons can be observed) — at a distance of 1.02 Å from the nitrogen (this distance was shown to lead to optimal data reproduction after adjusting the internuclear distance along the same vector).

Here, in a first step, the tensor determination was improved by allowing a `1D-GAF` motion during the tensor parameter determination, during which step the orientation of all planes were simultaneously optimized. Then a backbone structure was determined using a new approach: DYNAMIC-MECCANO. This approach is based on the MECCANO (Molecular Engineering Calculations using Coherent Association of Nonaveraged Orientations) method that was developed to sequentially determine protein structure using only RDCs [165]. The principle of this approach is to orient one after each other all the protein peptide planes according to experimental RDCs. Here this approach was reused but a dynamic averaging using `1D-GAF` motion was added to the description of the RDCs during the sequential structure determination. In order to remove the intrinsic two-fold RDC degeneracy of peptide planes, a harmonic potential was introduced to force the tetrahedral junction to an ideal geometry. Eventually a complete `3D-GAF` analysis was applied leading to very similar results to the previous analysis determined using the high resolution crystal structure. The backbone coordinates determined using this approach were remarkably similar (backbone RMSD of 0.5 Å) to the high resolution crystal structure. These approaches were tested through cross-validation of RDCs that were not used in the analysis (for example $^1D_{NH}$ couplings), demonstrating for the first time that a dynamic description better reproduced independent data than an optimally applied static approach.

Extensive simulations were eventually carried out, by comparing `3D-GAF` analysis and molecular dynamics trajectories, [166] in order to validate and estimate the accuracy of the approach. This resulted in the observation that the approach is robust when four sufficiently different alignment tensors are available, demonstrating one of the potential advantages of using a model that combines the orientational averaging properties of multiple RDCs in the same structural unit.

3.7 CONCLUSION

Dynamic studies of slow timescale motions have excited a great deal of interest over recent decades. Currently, a large range of approaches are

1 PDB codes and resolution (Å) of diverse structures: 1ejg (0.54), 1fy5 (0.81), 1gci (0.78), 1hje (0.75), 1m40 (0.85) 1pq7 (0.80), 1ssx (0.83), 1ucs (0,62) 3pyp (0.85), 1cq2 (2.00).

available, some numerical, such as the ones derived from MD simulations, some analytical such as model free approaches or the GAF model.

Even if some approaches such as the one based on the GDO, can be achieved with RDCs measured in a single alignment medium, most of the methods and especially the analytical methods require a good deal of experimental measurement. This issue can be crucial in the case of complex systems and this is why most of the presented studies are currently applied to model systems.

The structural information content of RDCs is often very high and therefore, omitting to incorporate dynamic averaging can give rise to a reasonable description of the experimental data. As models of dynamics are generally more complex than the static description, their validity needs to be tested and often statistical tests are essential to ensure their relevance.

A great deal of effort has been dedicated to the characterization of dynamics at slow timescales and the possibility to study internuclear vector dynamics on a site specific basis has been demonstrated. Nevertheless some points remain unresolved, such as the determination of the absolute level of dynamics in macromolecular systems. Most of the studies presented here require the scaling of the extracted order parameters against some reference. Usually the reference that is selected is the level of order determined from a model-free analysis of ^{15}N relaxation. This is a reasonable way to proceed but considering the dynamic information content of RDCs we were interested in determining the extent to which one could quantitatively determine protein dynamics from RDCs, and whether it is possible to avoid the use of other experimental techniques to determine the amounts of observable dynamics. This would be especially important as using sets of data sensitive to a clearly defined times window, e.g. only RDCs, will give further insight into the repartition of the dynamics over timescales in biological systems.

In many cases a structural input is required, for example to determine starting estimates for tensors properties. This can potentially introduce a bias in the determination of dynamics. The possibility of determining dynamics without prior structural information will allow an estimation of the dynamics present in the system and then, at least for folded systems, the ability to construct a structural model which incorporate this dynamic information.

4

QUANTITATIVE AND ABSOLUTE DETERMINATION OF BACKBONE MOTION IN UBIQUITIN

ABSTRACT

The determination of the absolute level of dynamics present in folded proteins on timescales up to the millisecond remains a challenging but important issue for the understanding of protein stability and function. A method, called SF-GAF, based only on the interpretation of RDCs, is developed to achieve this goal for protein backbone dynamics, in the absence of structural input or reference to external experimentally measured quantities. This analytical method, based on the GAF model and applied to Ubiquitin, allows the quantification of both direction and amplitudes of peptide plane motions. The relevance and robustness of the protocol is extensively tested and results are compared to complementary experimental and numerical approaches.

4.1 INTRODUCTION

As discussed in Chapter 3, characterization of motion on timescales that are longer than the rotational correlation time of the molecule remains a challenging obstacle that must be overcome before establishing the relationship between intrinsic protein dynamics and biological function. Remarkable progress has been made in recent years concerning the depiction of motional modes in proteins from extensive RDCs measurements, leading to the accurate characterization of the relative distribution of N_i-H_i^N order parameters along the protein backbone. The quantitative determination of dynamics remains challenging however, and as described, has often been resolved by referencing or scaling observed motional parameters, such as order parameters, to external references such as those derived from spin relaxation. This issue is particularly important, firstly because very little is known about the absolute level of dynamic fluctuations present in proteins, and secondly because RDCs are especially prone to errors of absolute amplitude of the

dynamics, due to the potential for a component of the dynamic averaging of RDCs to be absorbed into the estimated magnitudes of the alignment tensors. This is always a critical point for the relevance of order parameters derived from RDCs, as, considering equation 2.23 the similar dependence of the properties of the tensor and the true local dynamics defined by the second order averaged spherical harmonics, underlines the impossibility of accurately characterizing one aspect while disregarding the other. For these reasons the quantification of the amount of dynamics present at slow timescales remains an open question.

Compared to previous approaches describing 3D-GAF averaging of spherical harmonics to analyze RDCs, the methods presented here (see Sections 2.53 and 2.54) are computationally much more efficient, reducing the calculation time by a factor of ~50, and thereby providing access to previously inaccessible computational studies. The aim of this chapter, that focus on Ubiquitin (see Annexe B), is to use this increased efficiency to address the simple question: how precisely can we estimate the true level of dynamics present at timescales up to the millisecond in folded proteins?

4.2 MATERIALS AND METHODS

4.2.1 *Experimental Data*

Experimental RDCs emanating from a large RDC datasets have been previously compiled [68, 93, 153, 157, 159]. These data are summarized in a published SCRM analysis [153]. These RDC datasets are derived from 24 different alignment media, all containing $^1D_{NH}$ couplings, five of them $^1D_{C'H^N}$ and $^1D_{C'N}$ and two including $^1D_{C'C^\alpha}$. The data were selected from a larger ensemble using a SECONDA analysis [153] (see Section 3.2.2). The type of alignment and the kind of couplings are summarized in Table 1.

4.2.2 *Simulated Data*

Simulated RDCs for protein GB3 were obtained by averaging static RDCs over 1000 snapshots of a molecular dynamics simulation [166]. $^1D_{NH}$, $^1D_{C'C^\alpha}$ and $^1D_{C'N}$ were simulated for each peptide plane and noise was added using a Gaussian random distribution with standard deviation of 0.26 Hz for $^1D_{NH}$ and 0.10 Hz for $^1D_{C'C^\alpha}$ and $^1D_{C'N}$.

Table 1 – Alignment media used for the SF-GAF Ubiquitin analysis. All RDC datasets and their precise experimental characteristics can be found in the indicated references. The number of RDCs included in each dataset are indicated. Abbreviations are explicated above the table.

Nb	Medium	$^1D_{NH}$	$^1D_{C'N}$	$^1D_{C'H^N}$	$^1D_{C'C^\alpha}$
0	DMPC:DHPC [68]	57	61	61	58
1	DMPC:DHPC:CTAB [68]	57	63	63	53
2	PM 40 mM NaCl [159]	57			
3	PM 80 mM NaCl [159]	57			
4	PM 250 mM NaCl [159]	57			
5	PM 20 mM NaCl [159]	57			
6	Helfrich phase [159]	56			
7	Pf-1 phages [159]	57			
8	PEG/hexanol [159]	57			
9	PEG/hexanol [157]	57			
10	Pf1 phages [157]	54			
11	polyacrylamide gel [157]	54	63	60	
12	Helfrich phase [157]	55	64	64	
13	PEG/hexanol [153]	55	63	64	
14	Pf-1 phages 100 mM NaCl [153]	65			
15	positively charged gel [153]	61			
16	Pf1 phages/gels [93]	51			
17	Pf1 phages/gels [93]	54			
18	Pf1 phages/gels [93]	53			
19	Pf1 phages/gels [93]	59			
20	Pf1 phages	64			
21	DLPC:DHPC:SDS [153]	59			
22	DMPC:DHPC:C14PC [153]	64			
23	DMPC:DHPC:CTAB [153]	62			

DMPC: dimyristoyl-phosphatidylcholine
DHPC: dihexanoyl-phosphatidylcholine
C14PC: tetradecylphosphatidylcholine
CTAB: N-cetyl-N,N,N-trimethylammonium bromide
PM: Purple Membrane
PEG: *n*-alkyl-poly(ethylene glycol)

4.2.3 *Peptide Plane Geometry*

The definition of peptide plane geometry can be found in previous GAF analyses [167], see Section 3.6.3. Coordinates can be found in Table 2. The precise position of the amide proton was investigated by using diverse N_i-H_i^N bond lengths along the N_i-H_i^N vectorial direction. All analyses were carried out using 1.020 Å, in agreement with previous GAF analyses [167] and 1.024 Å as more recently proposed [97]. The analysis presented here uses 1.020 Å, the key results using 1.024 Å are summarized in Annexe C.3.

Table 2 – Two dimensional cartesian coordinates of all atoms constituting the idealized peptide plane used from GAF studies. This geometry corresponds to a N_i-H_i^N bond length of 1.020 Å.

Atoms	x	y
${}^{\alpha}C_{i-1}$	0.000	0.000
C'_{i-1}	1.430	−0.529
O_{i-1}	1.669	−1.746
N_i	2.404	0.382
H_i^N	2.161	1.373
${}^{\alpha}C_i$	3.806	−0.006

4.2.4 *Target Function and Minimization Algorithm*

The target function used during the optimization procedure is a standard χ^2 function, expressed as:

$$\chi^2 = \sum_{r,i,m} \left(\frac{D_{r,i,m}^{calc} - D_{r,i,m}^{exp}}{\delta_{i,m}} \right)^2 \tag{4.1}$$

where r runs over the different peptide planes, i over the type of RDCs (e.g. ${}^1D_{NH}$), and m over the alignment media.

It is worth emphasizing that this target function depends on the mean orientation $(\theta, \phi)_{r,i}$ of each studied internuclear vector, on their dynamics (if invoked) — $(\sigma_\alpha, \sigma_\beta, \sigma_\gamma)_r$ for GAF motions, S_r for isotropic diffusion in a cone — and on the alignment tensor characteristics $(A_a, A_r, \alpha, \beta, \gamma)_m$. Here the orientation $(\theta, \phi)_{r,i}$ is characterized with a subscript i as the different vectors of each peptide plane react differently to the different motional modes. Due to the use of a perfectly rigid peptide plane, knowing the orientation of the plane is enough to deduce all in-plane vector orientations. Therefore the different orientations of all in-plane vectors can be deduced

from three angles that define the plane orientation and the idealized peptide plane topology (see above).

An alternative target function [168] was also used, which behaves as a standard χ^2 for deviations smaller than a few standard deviations and which reaches a plateau value for larger deviations, and whose analytical expression is:

$$\xi^2 = c^2 \sum_{r,i,m} \left[1 - exp \left[- \left(\frac{D^{calc}_{r,i,m} - D^{exp}_{r,i,m}}{c\, \delta_{i,m}} \right)^2 \right] \right] \tag{4.2}$$

with c an adjustable parameter that will determine the behavior of the target function against outliers. Here a classical value of $c = 2.9846$ was used. The whole analysis was carried out with this function, results were found to be similar to the standard χ^2 approach, but the combination of the removal of outliers (see below) and the standard χ^2 target function were found to be slightly more efficient and robust, compared to the ξ^2 function. Thus results presented here use the χ^2 protocol.

All minimizations, presented here or in later chapters, were performed using in-house software. Extensive calculations involving a large number of degrees of freedom and lengthy minimization protocols can easily become time consuming. The minimization procedures are commonly performed using a two step protocol:

1. The first minimization is performed using the Particle Swarm Optimisation (PSO) algorithm [169]. This is a stochastic algorithm, which presents some similarities to better known genetic algorithms (see Section 10.2.4), was developed by Eberhart and Kennedy and is inspired by the social behavior of bird flocking. Minimization can be described by considering the motion of a population in the solution space. The evolution is performed iteratively, but differently to genetic algorithms, the evolution is not achieved by an evolution operator such as reproduction, but by modifying the position and velocity of the different members of the population. Each member of the population is directed towards its personal minimum and the global minimum already found, using random acceleration towards these two points. Starting parameters are selected at random. PSO optimizers were selected for this application because of their high efficiency in high-dimensional solution space [170].

2. In a second step a deterministic algorithm such as Broyden-Fletcher-Goldfarb-Shanno (BFGS) protocol (which is a quasi Newton method)

[171] is used to accurately determine the function minimum from the end point of the previous step.

The first step minimization is based on a stochastic algorithm. In order to avoid inappropriate convergence due to a particular ensemble of randomly generated numbers, the minimization is repeated until clear reproducibility is achieved. Using this minimization approach the computation time necessary to optimize the orientation and motional amplitude for a peptide plane undergoing `3D-GAF` motion, with known alignment tensors, is on the order of one minute.

4.2.5 *Static and Dynamic Models*

In this analysis four different models are used to analyze the RDCs:

- Static model: only structural features $(\theta, \phi)_{r,i}$ are determined, no dynamics are invoked.

- `S` model: the model optimizes the orientation $(\theta, \phi)_{r,i}$ and an order parameter S_r for each peptide plane. This motional description corresponds to the diffusion in a cone model described in Section 2.6.2.

- `1D-GAF/S` model: the model optimizes the orientation $(\theta, \phi)_{r,i}$ and a single dynamic amplitude parameter for each peptide plane. To determine the dynamic parameter the analysis is repeated four times, once using a diffusion in a cone model and optimizing an order parameter S_r, three times with a `1D-GAF` model in order to sample successively the three possible reorientational models by optimizing $\sigma_{\alpha,r}$, $\sigma_{\beta,r}$ or $\sigma_{\gamma,r}$. The model that gives the best reproduction of the data is retained.

- `3D-GAF` model: the model simultaneously optimizes for each peptide plane the vector orientations $(\theta, \phi)_{r,i}$ and the three amplitudes of reorientation $(\sigma_\alpha, \sigma_\beta, \sigma_\gamma)_r$.

4.2.6 *Alignment Tensors, Weight Determination and Outliers Detection*

1. In all studied alignment media, the relative weight of each RDC type is initially set at 10% of the range of the experimental values.

2. The alignment tensors are determined using a completely structure free approach. First of all, a rough estimation of all static tensors is obtained using a static description. Here only peptide planes having extensive RDC datasets (more than 35 RDCs) are considered. The

approach simultaneously optimizes the five tensor components (magnitude, rhombicity and relative orientations) and the orientation of each peptide plane. In order to deal with the large solution space (many degrees of freedom), the optimization starts with an arbitrary fragment of the sequence whose length is increased until all of the protein is included in the minimization process. The starting window was set in different positions of the protein and with different fragment lengths in order to test for local bias: good convergence was found, and after dynamic optimization (see subsequent steps) the potential influence of the starting fragment was found to be entirely negligible.

3. Based on statistical analysis of the dispersion of the reproduction of experimental data, the weights $\delta_{i,m}$ of the different RDCs are fixed in order to avoid over-fitting of a particular data type. A different weight is used for each kind of RDCs in each different alignment medium. Median or median absolute deviation are commonly used to estimate this kind of dispersion. Nevertheless scale estimators such as S_n and Q_n were shown to be slightly more robust [172]. Here weights are fixed according to the Q_n scale estimator. During this step, potential outliers can be detected and removed to improve harmonic fitting of the data, if deviation between experimental and calculated RDCs remains higher than 6 Q_n values. An alternative to removal of outliers is to use a modified target function (see above).

4. The properties of the dynamic tensors are determined during a second fitting procedure. Starting from static tensors and re-weighted RDC datasets, peptide planes containing a sufficiently high number of RDCs (still more than 35) are oriented and dynamically characterized in a sequential manner, according to a `1D-GAF/S` analysis. Then the five components of the tensors are optimized. This protocol is applied iteratively in order to obtain a stable solution.

5. A new weighting step, similar to step 3 was carried out, according to `1D-GAF/S` data reproduction.

6. In order to systematically estimate the absolute alignment amplitude, an overall scaling factor K_A is applied to all tensors (A_a and A_r). According to `S`, `1D-GAF/S` and `3D-GAF` motional models, a complete analysis is performed for each value of K_A, varying in the range of 0.900 to 1.100 in steps of 0.002 (where the value $K_A = 1$ corresponds to the optimal tensors found during the previous optimization, with `1D-GAF/S` model of motion). Two kinds of datasets were used: the complete dataset with all RDCs, and ten different sub-ensembles where two $^1D_{NH}$ RDCs are randomly removed per peptide plane containing

more than 20 $^1D_{NH}$ couplings. Data reproductions obtained for the ten different calculations, with the reduced datasets, were averaged in order to decrease the impact of selecting a particular sub-ensemble of RDCs. The optimal K_A value, corresponding to the best reproduction, according to `3D-GAF` model for unused couplings, is applied to all tensors.

4.2.7 *Local Dynamic Study*

This analysis, which uses all tensors and weighting determined in the previous step, is made using the following steps:

1. Using the `3D-GAF` approach, the orientation $(\theta, \phi)_{r,i}$ and dynamic parameters $(\sigma_\alpha, \sigma_\beta, \sigma_\gamma)_r$ are determined sequentially for each peptide plane. This model is called `M-III`, as the three amplitudes of reorientation are optimized.

2. An averaged value of $(\sigma_\alpha)_r$ diffusion angles over all peptide plane $\sigma_{\alpha,av}$ is estimated in a `3D-GAF` analysis where all peptide planes share the same reorientation amplitude $\sigma_{\alpha,av}$. Here $(\sigma_\beta, \sigma_\gamma)_r$ for all peptide planes and $\sigma_{\alpha,av}$ are simultaneously optimized (orientations are fixed to the one obtained during the previous step).

3. A similar analysis is applied using the two averaged angles, the previously determined $\sigma_{\alpha,av}$ and $\sigma_{\beta,av}$, the only locally optimized motional amplitude being $(\sigma_\gamma)_r$.

4. A `3D-GAF` analysis is applied, plane by plane, using $\sigma_{\alpha,av}$ and $\sigma_{\beta,av}$ averaged values. During this minimization the peptide plane orientation and $\sigma_{\gamma,r}$ are optimized. This model of motion is called `M-I`.

5. The same analysis is applied using only $\sigma_{\alpha,av}$. In this model, named `M-II`, the peptide plane orientation and $(\sigma_\beta, \sigma_\gamma)_r$ are optimized.

6. According to AIC or F-test analysis for each peptide plane the most appropriate model between `M-I`, `M-II` and `M-III` is selected. This procedure avoids over-fitting.

As models are nested, i.e. increasing the complexity of a model corresponds to using the previous model with supplementary terms, both AIC or F-tests are suitable.

The AIC (standing for Akaike's information criterion) [173] is a way to compare models using this AIC value that indicates which model is more

likely to be correct. For models where the quality of the fit is estimated through a maximum of likelihood function (e.g. standard χ^2), a modified AIC, the AIC_{χ^2} can be used to estimate the quality of the model[1]. If the model fits exactly the same number of points the AIC_{χ^2} can be expressed as [174]:

$$AIC_{\chi^2} = \chi^2 + 2k \tag{4.3}$$

where k is the number of parameters of the model. This represents a way to select models through Occam's razor principle as it balances the quality of data reproduction and the complexity of the model. Comparing two models, the one with lowest AIC or modified AIC will be more likely to be correct.

The F-test [173] will compare two models 1 and 2. If model 1 is nested in model 2, the increase of the number of parameters from 1 to 2 should be followed by a decrease in χ^2 ($\chi^2_2 < \chi^2_1$). The F-test will estimate whether improvement in data reproduction is just obtained by chance. The F value to compare two models is obtained by:

$$F = \frac{\chi^2_1 - \chi^2_2}{\chi^2_2} \frac{k_2}{k_1 - k_2} \tag{4.4}$$

This value can be compared to tabular, analytically calculated or Monte-Carlo simulated values, to determine whether the obtained value is inferior to the P value corresponding to the level of desired confidence often 5% (or 10%). If the obtained F is smaller than the P value of 5%, it means that, considering the simplest model as correct, the probability to observe the improvement obtained by using the more complex model by chance is less than 5%.

4.2.8 *Accuracy Estimation*

The accuracy of the method was probed using Monte-Carlo simulations. Noise-based Monte-Carlo simulation was also used to test for statistical significance. For each of the RDCs, noise was added using a Gaussian noise distribution of standard deviation $\delta_{i,m}$, the weight estimated during the analysis, in order to obtain a new simulated RDC dataset. 1000 different datasets were produced in this way. For each, a complete local dynamic analysis was applied, resulting in a distribution of calculated values. The distribution of models of motion (`M-I`, `M-II` and `M-III`) corresponds to the distribution obtained during local dynamic analysis. The standard

1 Other modified AICs exist such as AIC_c. The most appropriate criterion depends on the considered problem.

deviation of these distributions was interpreted as an estimation of the absolute accuracy of the method. For models `M-I` and `M-II` the some amplitudes of reorientation are fixed to an averaged value. Their accuracy was thus estimated using Monte-Carlo simulations realized according to a standard `3D-GAF` model.

4.2.9 *Cross-Validations*

The local dynamic analysis was also tested through cross-validations. In such an analysis a set of RDCs is removed from the active dataset and are called passive data. Using active data, a complete local dynamic analysis is applied. The resulting dynamic and orientational information are used to back-calculate passive RDCs which are then compared to experimental data.

Cross-validation using a static model was applied in a parallel analysis. In this analysis, the same passive set of RDCs was used. The tensors used were those optimized for a static analysis.

4.3 RESULTS AND DISCUSSION

4.3.1 *Absolute and Quantitative Determination of the Alignment Tensors*

The tensor determination protocol was applied to the Ubiquitin experimental dataset. The results are shown in Table 3. In tensor determinations carried out in previous analyses, the accuracy of relative magnitude, rhombicity and orientation was shown to be accurate [104], and previous `GAF` studies applied to simulated data [166] demonstrated that similar a `1D-GAF` tensor determination correctly reproduced these quantities. Nevertheless, some underestimation of the absolute alignment amplitude was always detected. Thus steps 1 to 4 above should give a relative good tensor properties definition, but the significance of the precision of the magnitude merits further investigation.

In order to characterize more precisely the question of absolute amplitude of the motion, the repetitive protocol involving the overall scaling factor K_A was applied. Results are shown in Figure 16. The same protocol was applied to simulated data on protein GB3 for comparison. For GB3 the value of $K_A = 1$ correspond to the amplitudes of the tensors used to simulate the data, and the results for direct analysis using an `S`, `1D-GAF/S` or `3D-GAF` are shown in Figure 17. The general features of the results are highly similar

Table 3 – Alignment tensors determined during Ubiquitin analysis. Axial and rhombic components are those obtained after the complete analysis, i.e. after application of the experimentally determined 1.02 scaling factor.

Tensor	A_a (10^{-4})	A_r (10^{-4})	α (°)	β(°)	γ(°)
0	9.12	1.25	122.51	88.04	−91.36
1	15.80	8.15	−155.27	85.50	89.71
2	4.20	0.13	79.29	139.42	138.87
3	3.47	0.11	79.28	138.66	138.75
4	2.98	0.13	83.44	138.26	139.42
5	3.44	0.42	68.52	134.88	141.44
6	3.90	0.11	91.90	139.66	139.82
7	−10.16	−2.78	−54.49	27.12	−14.03
8	−3.41	−1.06	4.86	43.65	−33.20
9	8.52	4.14	−163.92	72.69	96.63
10	5.15	2.24	−160.06	74.88	94.98
11	13.42	7.23	−143.66	81.42	88.62
12	6.79	0.51	75.98	139.28	138.83
13	−7.68	−2.53	−51.60	25.86	−13.99
14	11.57	5.46	−164.56	73.93	97.35
15	8.41	3.88	88.98	156.04	179.62
16	−13.02	−3.11	−39.82	47.46	−20.18
17	−10.01	−2.15	−48.30	48.17	−19.86
18	−10.45	−1.77	−52.61	44.31	−18.55
19	6.76	3.57	−138.53	81.03	88.50
20	−3.41	−1.06	4.87	43.66	−33.19
21	20.77	3.39	20.99	100.71	81.06
22	8.58	1.24	−138.86	90.65	89.42
23	18.98	9.63	−156.00	86.14	89.66

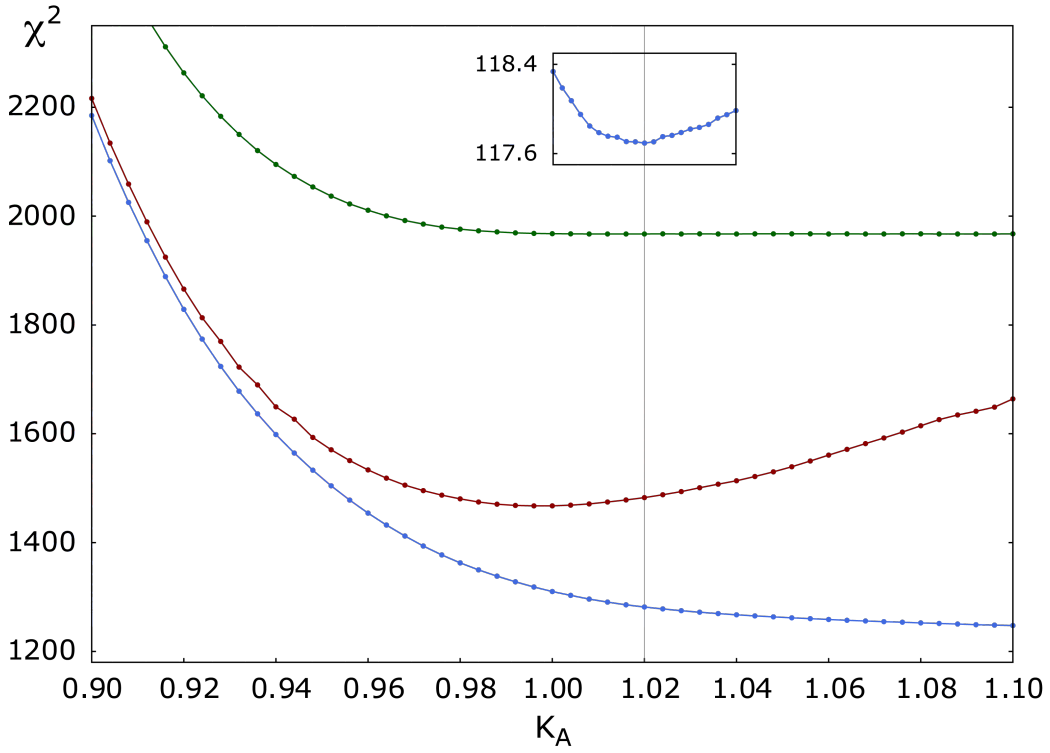

Figure 16 – Ubiquitin experimental data. Effect of the alignment tensor scaling on direct data reproduction χ^2 according to model `S` (green), `1D-GAF` (red), `3D-GAF` (blue). The scaling is applied according to K_A. The value $K_A = 1$ corresponds to the tensors obtained after `1D-GAF` optimization. The inset corresponds to indirect data reproduction according to `3D-GAF`. Gray line corresponds to the optimal tensors.

allowing us to suppose that the behavior of the analysis of simulated data for GB3 remains valid for the Ubiquitin analysis.

Starting from very low alignment tensor eigenvalues, i.e. small K_A, the three models show a decrease of the overall χ^2 as K_A increases, moreover the χ^2 decreases by changing the model from `S` to `1D-GAF` and finally `3D-GAF`. The first of these observations is due to the fact that too small tensors do not allow for any dynamics, and therefore, by decreasing the tensor magnitude, the descriptions converge to a static situation. The second observation corresponds to a manifestation of the increasing number of degrees of freedom between the different models.

By increasing K_A value, different phenomena can be observed. Concerning the `S` model, the χ^2 rapidly reaches a plateau value. For the `1D-GAF/S` description, the shape of the χ^2 curve is quite different, exhibiting a minimum, which corresponds to a data reproduction significantly better than the one obtained using the `S` model. Eventually the `3D-GAF` model, which always

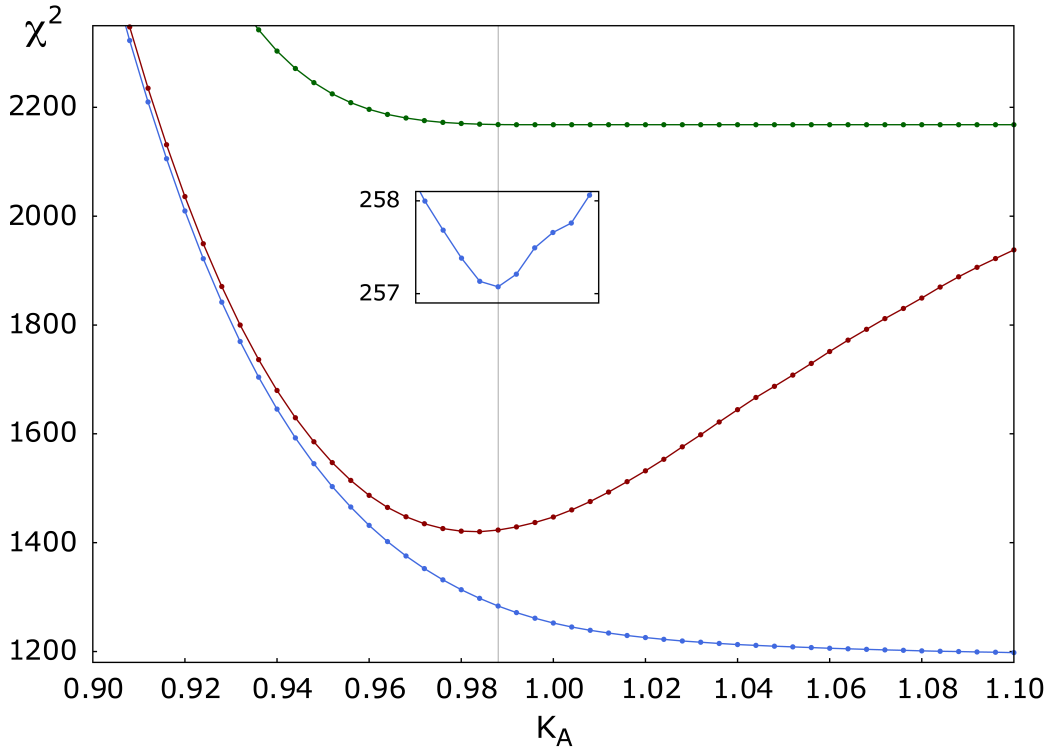

Figure 17 – GB3 simulated data. Effect of the alignment tensor scaling on direct data reproduction χ^2 according to model `S` (green), `1D-GAF` (red), `3D-GAF` (blue). The scaling is applied according to K_A. The value $K_A = 1$ corresponds to the tensors used for simulation. The inset corresponds to indirect data reproduction according to `3D-GAF`. Gray line corresponds to the optimal tensors.

gives the best data reproduction, exhibits a decreasing χ^2 curve, whose curvature decreases with increasing K_A.

Data reproduction according to the `S` description indicates that isotropic motions are not able to properly determine the tensor magnitude, as changing of the overall scaling K_A leaves the quality of data reproduction unchanged. A comparison of data reproduction between the `S` and `1D-GAF/S` models underlines the fact that anisotropy is required in order to correctly describe the peptide plane motion, as the only difference between the two models is that in the `1D-GAF` model isotropic motion can be exchanged for an anisotropic model if the data are better reproduced by this anisotropic description. The minimum, which corresponds to the results of the previous analysis (step 4), demonstrates that this highly anisotropic description of the motion correctly determines the tensor magnitudes.

`3D-GAF` χ^2 evolution shows that this motional description always best reproduces the data, which is coherent with the fact that this model has the most adjustable parameters. Even at high K_A values the `3D-GAF` model is able to fit the data. For the `3D-GAF` model an increase of the K_A, can be seen

as a situation where the system undergoes its "true" dynamics (i.e. the one from the MD trajectory) and an additional isotropic motion (which further scales as the tensor values increase). By simultaneously increasing the three amplitudes of reorientation, this excess of dynamics can be accommodated by the `3D-GAF` motion. This situation can lead to an improvement of the data reproduction which does not have any physical meaning. This is clearly the case for GB3, where the data were simulated: a decrease in the χ^2 when K_A exceeds 1 corresponds to detection of more motion than was actually present in the MD trajectory.

In order to overcome the issue of incorrectly estimating tensor amplitudes during the analysis, an indirect analysis was applied. Concerning `S` and `1D-GAF/S`, the results are essentially the same as in the direct analysis and therefore data are not shown. For the `3D-GAF`, a minimum can be found using indirect data reproduction as shown in the insets of Figures 16 and 17. The presence of this minimum can be rationalized as follows:

- Firstly, increasing the three reorientation amplitudes does not exactly correspond to an isotropic motion. This has been verified by simulating RDCs dynamically averaged through a perfect `3D-GAF` motion: the same analysis, with artificial tensor scaling, gives rise to a shallow minimum at the K_A value corresponding to the tensor used for data simulation using a direct analysis. Nevertheless, the shallowness of this minimum suggests that it can be easily missed in the presence of noise. An indirect approach can however reveal its presence.

- Secondly, the indirect approach is sensitive to over-fitting. In fact, if the model starts to fit the noise content of a dataset, it will make the direct χ^2 of the analysis decrease but not the indirect χ^2.

In the case of GB3 the indirect analysis using the `3D-GAF` model allows the estimation of the tensor magnitudes with an accuracy close to 1%. By analogy the corresponding minimum found in the Ubiquitin dataset analysis is then assumed to give a quantitative estimation of the alignment tensors.

4.3.2 *Local Dynamics*

For the local dynamic study of Ubiquitin, the two averaged orientation angles were optimized to $\sigma_{\alpha,av} = 4.26\,^\circ$ and $\sigma_{\beta,av} = 9.24\,^\circ$. Optimization was applied for each peptide plane, but results were considered to be robust enough when more than 10 RDCs per peptide plane were available. Thus complete characterization was applied to 63 peptide planes. Four planes (22, 31, 72 and 74) were probed with fewer than 20 couplings: their analysis

is expected to be less robust. The optimized parameters (corresponding order parameters and amplitudes of motions) can be found in Tables 17 and 18, in Annexe C. N_i-H_i^N order parameters, σ_γ and σ_β results are shown in Figure 18.

The model selected for the dynamics was `M-I` for 22 peptides planes, `M-II` for 30 planes and `M-III` for the 11 others. The different models are more or less distributed homogeneously along the sequence. The only notable point is that residues in α-helix very rarely exhibit `M-III` (a single residue presents this configuration).

After model selection and amplitude optimization, an averaged value of the three reorientation angles can be calculated, leading to:

$$\langle \sigma_\alpha \rangle = 7.01\,^\circ, \quad \langle \sigma_\beta \rangle = 8.49\,^\circ \quad \text{and} \quad \langle \sigma_\gamma \rangle = 12.50\,^\circ \tag{4.5}$$

These values were found to be in a similar range to that found in experimental `3D-GAF` of GB3 studies, with slightly higher values for $\langle \sigma_\alpha \rangle$ and $\langle \sigma_\beta \rangle$ [164, 167].

The occurrence of a reorientation amplitude of zero was observed for some residues. This phenomenon was also observed in previous `GAF` studies [164, 166, 167]. Considering the three reorientations (σ_α, σ_β, σ_γ), the percentage of zero reorientation angles is (0%, 16%, 13%), whereas in previous GB3 studies it was (49%, 26%, 15%). Zero γ-motions seem unrealistic, especially for planes flanked by neighbours with $\sigma_\gamma \sim 15\,^\circ$. In addition, the distribution of zero σ_β or σ_γ does not match with any specific amino-acid or with some residue involved in interaction such as hydrogen bonding. The results seem to be mainly due to the lack of accuracy of the `GAF` model to determine small reorientation angles (smaller than around $\sim 5\,^\circ$) [163], but can also be due to small inconsistencies in the dataset (such as experimental noise...), lack of data (for planes 22 and 31) or to the inability of the model to fit a particular motion. Results are shown for these peptides planes, but caution has to be taken for their precise interpretation.

Results in term of order parameters S^2_{NH} can be visualized on the Ubiquitin structure (PDB code 1d3z [175]) in Figure 19.

4.3.3 *Comparison with ^{15}N Relaxation*

^{15}N relaxation order parameters were provided by Brüschweiler and coworkers [149]. A comparison of order parameters between ^{15}N relaxation $S^2_{NH,REL}$ and those derived from `SF-GAF` analysis $S^2_{NH,GAF}$ is shown in Figure 18. Globally, order parameters $S^2_{NH,REL}$ and $S^2_{NH,GAF}$ show similar profiles in the

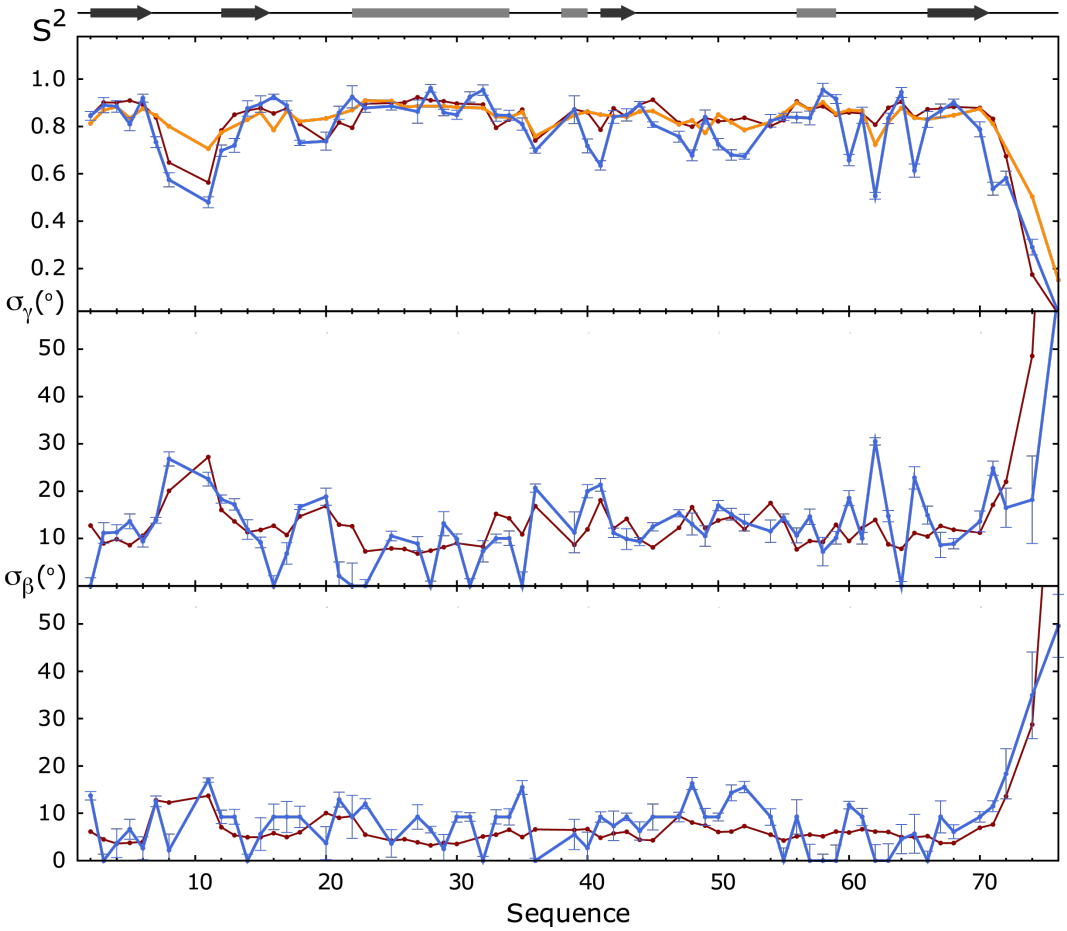

Figure 18 – Local Ubiquitin dynamics obtained through `SF-GAF` analysis. N_i-H_i^N order parameters (upper panel) and amplitudes of reorientations for γ-motion (central panel) and β-motion (lower panel) derived from `SF-GAF` analysis (blue), 400 ns MD simulation (red) or ^{15}N relaxation (orange). Secondary structures are indicated on the top of the figure. Grey boxes indicate α-helix and darker arrows indicate β-sheet.

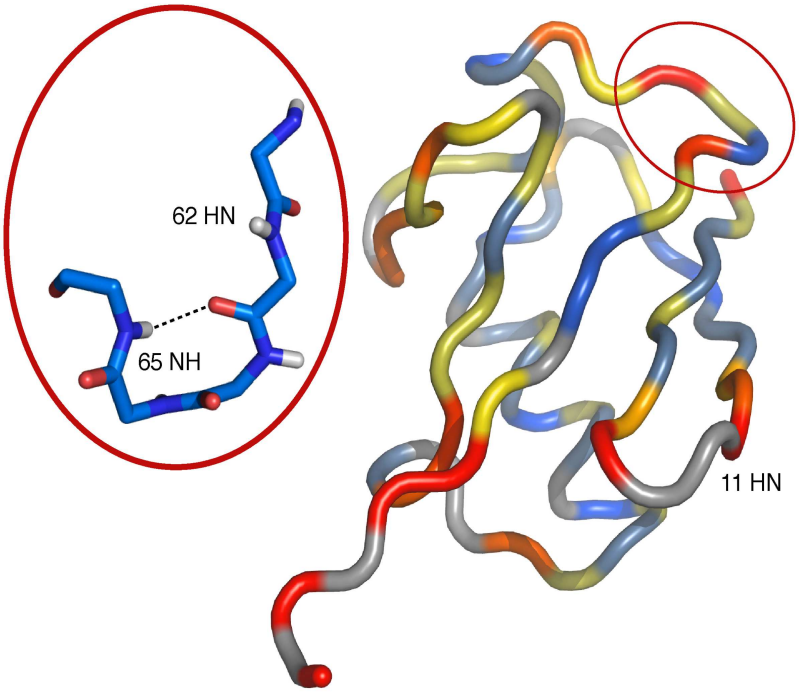

Figure 19 – SF-GAF N_i-H_i^N order parameter ($S^2_{NH,GAF}$) of Ubiquitin represented on 1d3z structure. Color scale from dark blue ($S^2_{NH,GAF} = 1.0$) to dark red ($S^2_{NH,GAF} = 0.5$) via green, yellow, orange. Grey: not determined. Insert: Turn region 62-65 showing high γ- and α-motions amplitude and the thermolabile hydrogen bond across this turn.

secondary structured regions, while the SF-GAF description leads to more dynamics in loop regions. Assuming uncertainties of 0.03 for $S^2_{NH,REL}$, for three peptide planes (16, 27 and 32) the $S^2_{NH,GAF}$ was found to be higher than the corresponding $S^2_{NH,REL}$. RDCs derived order parameters should not exhibit less dynamics than order parameters derived from spin relaxation as they are sensitive to longer timescales (see Chapter 2). Nevertheless the following points could lead to such a situation:

- The relaxation data are also prone to analytical errors. The constant value used for the ^{15}N CSA can be one of the potential source of error.

- Some of the problematic residues exhibit zero γ-motion ($\sigma_\gamma = 0$). As previously discussed this is potentially a non physical situation. A more accurate determination of the angle of reorientation could overcome this problem.

- The dynamics characterized by the SF-GAF approach concern the entire peptide plane, whereas ^{15}N relaxation focuses only on N_i-H^N_i internuclear vector. Therefore, if some supplementary motions occur for the N_i-H^N_i vector, which are not shared with the rest of the plane, an averaged dynamics will be determined in the GAF analysis, leading to the theoretical possibility of observing $S^2_{NH,GAF}$ higher than the $S^2_{NH,REL}$. In other words, the S^2_{NH} compared in the two approaches does not refer to identical dynamics: $S^2_{NH,REL}$ characterizes the N_i-H^N_i vector alone, whereas the $S^2_{NH,GAF}$ corresponds to the dynamics of the peptide plane, in the N_i-H^N_i direction. It is worth nothing that part of this possible differential dynamics can be absorbed in an effective N_i-H^N_i bond length. The results presented here with two different lengths nevertheless give highly similar results (see Annexe C). The existence of such motions does not seem to be general, considering the accuracy with which the GAF model can reproduce experimental and simulated data. Nevertheless, the site-specific situation may have to be considered.

A previous 3D-GAF analysis of seven relaxation rates measured in Ubiquitin has been published [149]. Due to the small amount of experimental data and the need to determine local and global correlation times associated to the movement, as well as the use of common descriptions for the carbonyl CSA tensors, (see Figure 20) only qualitative comparisons can be made. This study exhibits higher mobility for the γ-motion and a flatter distribution of amplitudes of motion which is coherent with the more narrow distribution of S^2_{NH} observed for ^{15}N relaxation.

The comparison of the distributions of $S^2_{NH,GAF}$ and $S^2_{NH,REL}$ is consistent with an overall motional picture where secondary structures undergo mainly

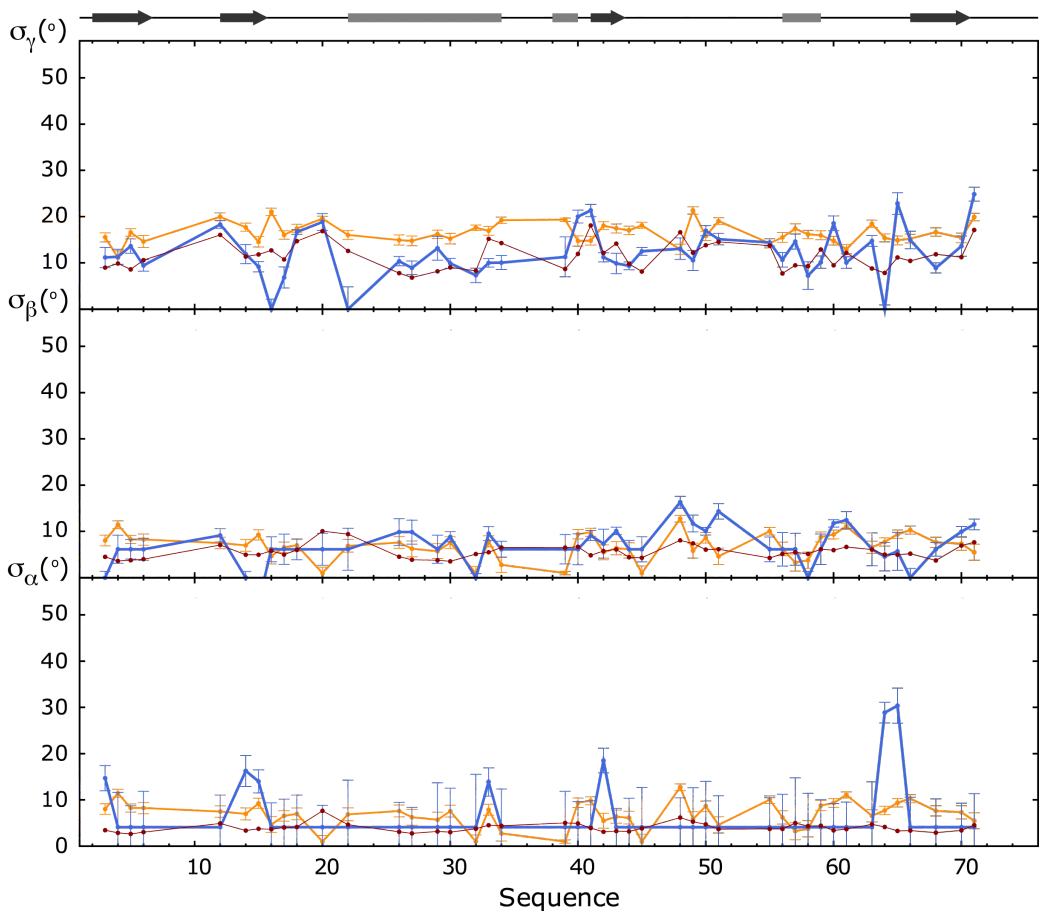

Figure 20 – Amplitudes of local reorientations in Ubiquitin through SF-GAF analysis. Amplitudes of reorientations for γ-motion (upper panel), β-motion (central panel) and α-motion (lower panel) derived from SF-GAF analysis (blue), 400 ns MD simulation (red) or GAF analysis of ^{15}N relaxation rates (orange). Secondary structures are indicated on the top of the figure. Grey boxes indicate α-helix and darker arrows indicate β-sheet.

fast dynamics, probed by both ^{15}N relaxation and RDCs measurements, whereas slower motions appear in less structured regions.

4.3.4 *Comparison with Molecular Dynamics Simulations*

The comparison with MD simulations potentially provides atomic detail of the nature of slower motions. The 400 ns MD trajectory used here was performed and analyzed by Brüschweiler and co-workers [148]. In order to extract amplitudes of motion, the trajectories were analyzed by diagonalizing a matrix formed by averaging over the trajectory orientations of the three vectors generating a GAF frame (e.g. the three orthonormal vectors of the $\mathcal{R}_\gamma$)[41]. This provides three reorientational amplitudes around the three orthogonal axes for each peptide plane. In this analysis, the reori-

entational amplitudes are sorted by decreasing value, corresponding to defining γ-motion as the largest motion. These three axes do not have to match with the three SF-GAF axes $\mathcal{A}_\alpha, \mathcal{A}_\beta$ and $\mathcal{A}_\gamma$, but there is an observed correspondence between the two sets of axes because the γ-motion is the dominant motion in the molecular dynamics simulation as it is the least constrained motion.

The amplitudes of motion and the S^2_{NH} of the two approaches can be seen in Figure 18. Comparison of motional amplitudes revealed some similarity between the SF-GAF and MD $(\sigma_\alpha, \sigma_\beta, \sigma_\gamma)$ distributions. Specifically, the distribution of σ_α was almost identically flat at the same $\sim 4^\circ$ value for both approaches, the SF-GAF exhibiting more dynamics only when the model M-III was selected. For $(\sigma_\beta, \sigma_\gamma)$, the distribution exhibits a larger range. Excluding the apparently ill-behaved residues with zero γ-motion, the profiles of amplitudes of reorientation are similar. On average, SF-GAF analysis shows a slightly higher amount of dynamics, and we note that, if the γ-motion exhibits less dynamics than the MD trajectory, a higher β-motion is observed. This may be due to the way by which amplitudes of motions are extracted from the two approaches: β- and γ-motions are along fixed axes in the SF-GAF, whereas they correspond to the two main reorientation directions in the MD approach. The major discrepancy came from residues 62 and 65 where a significantly higher γ-motion was found using the SF-GAF approach. Interestingly, residues 64 and 65 exhibit the two highest α-motions, see Figures 19 and 20.

This reorientation amplitude profile therefore results in slightly lower order parameters for the SF-GAF compared to the MD approach in both loops and secondary structures. It is worth noting that the $S^2_{NH,MD}$ are sometimes a bit higher than the $S^2_{NH,REL}$. The major discrepancy remains in the 62-65 loop region.

The generally good agreement between the two approaches further supports the SF-GAF approach, and suggests that much of the motion present in Ubiquitin occurs on timescales within the range of this MD trajectory. The discrepancy in the loop 62-65 could be explained by a slower motion.

4.3.5 *Comparison with the SCRM Approach*

The SCRM approach (see Section 3.5.1) was used to study a similar dataset of RDCs for Ubiquitin but which contained only the $^1D_{NH}$ couplings [153]. This analysis is model-free and thus does not contain the potential bias arising from the use of a physical model, but as applied by the authors, it requires an overall scaling in order to give results that can be compared with

$S^2_{NH,REL}$, as the alignment tensors are determined using a static approach [153]. Here, to facilitate comparison, $S^2_{NH,SCRM}$ were scaled in order to verify, for each $^1D_{NH}$ internuclear vector, $S^2_{NH,SCRM} \leqslant S^2_{NH,REL}$. Different $S^2_{NH,REL}$ are available in the literature for Ubiquitin. In the original SCRM publication [153] the authors used the ones obtained by Chang and Tjandra [176]. Those order parameters are smaller than the one presented here [149]. The choice of $S^2_{NH,REL}$ is of a great importance for the absolute amount of dynamics revealed by the SCRM approach. During a relaxation based study of Ubiquitin in interaction with an SH3 domain (see Chapter 8), we had the occasion to remeasure Ubiquitin ^{15}N relaxation. Our analysis gives good convergence with the one coming from Brüschweiler and co-workers (see Figure 91 in Annexe C), thus these S^2_{NH} will be used for scaling the SCRM. Comparison of $S^2_{NH,SCRM}$ and $S^2_{NH,GAF}$ are shown in Figure 21, original values of SCRM are also presented.

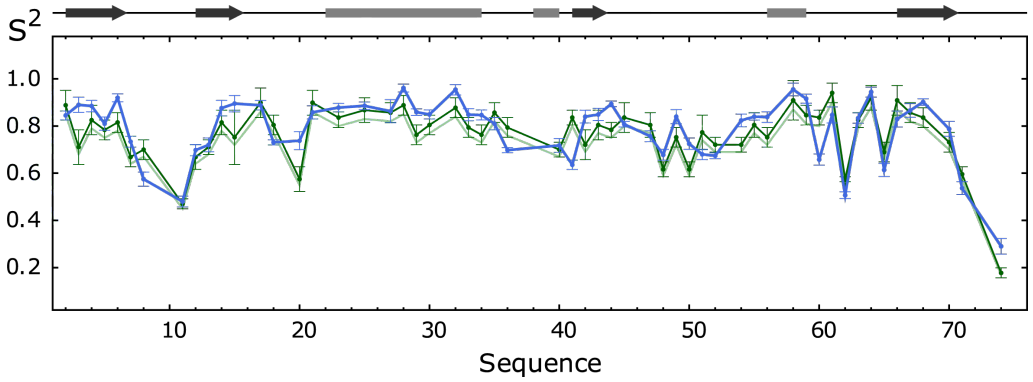

Figure 21 – Comparison between SF-GAF and SCRM derived N_i-H^N_i order parameters in Ubiquitin. SF-GAF results are in blue, the SCRM in green. SCRM order parameters were scaled to be equal to or lower than the relaxation derived order parameters [149]. Original SCRM values are presented in light green. Grey boxes indicate α-helix and darker arrows indicate β-sheet.

The obtained comparison shows very good convergence of the two approaches. No direct information can be obtained from the comparison of the amplitudes, due to the scaling of $S^2_{NH,SCRM}$. Nevertheless, scaling the order parameters from the SCRM approach as presented here results in a convergent level of slow dynamics for SCRM and SF-GAF. The agreement between the two methods in terms of site specific distribution of S^2_{NH} is very good, leading to essentially similar patterns. A few differences remain that may be due to the different hypotheses underlying the two methods or to the different information content of the two methods, as unlike SF-GAF, SCRM approach only takes into account $^1D_{NH}$. This very good agreement in S^2_{NH} relative distribution can be seen as an *a posteriori* validation of the hypothesis underlying GAF description of peptide plane motion: such a convergence between a model-based and a model-free approach cannot be expected if the hypotheses underlying the biophysical model were inappropriate.

4.3.6 *Robustness of the Approach*

One advantage of the GAF approach is that, requiring less data, by coupling vectors in the same structural unit, frees data for the purposes of cross-validation. The robustness of the approach was therefore tested with 34 independent cross-validation analyses, using SF-GAF and static models. The first 24 of these, corresponding to the successive removal of each of the different alignment media, the last 10 to the random withdrawal of 2 $^1D_{NH}$ per peptide plane.

In the first series of cross-validation analyses, all of the different cross-validations led to lower χ^2_{GAF} compared to χ^2_{STAT} (see Table 4). The averaged reduced χ^2 over the 24 alignment media is 4.2 for the static approach and 1.1 for the 3D-GAF one. The second series led to similar results, as shown in Figure 22, with average reduced $\overline{\chi}^2$ ($\overline{\chi}^2 = \chi^2/N$, where N is the number of considered RDCs) over the 10 randomly generated datasets of 3.7 for the static approach and 1.0 for the 3D-GAF method.

Table 4 – Ubiquitin SF-GAF cross-validations. Data reproduction quality expressed through reduced χ^2 for both SF-GAF ($\overline{\chi}^2_{GAF}$) and static ($\overline{\chi}^2_{STAT}$) approaches.

Tensor	$\overline{\chi}^2_{GAF}$	$\overline{\chi}^2_{STAT}$	Tensor	$\overline{\chi}^2_{GAF}$	$\overline{\chi}^2_{STAT}$
0	1.71	6.37	12	1.21	2.15
1	2.25	3.07	13	1.24	1.64
2	0.68	6.63	14	0.75	3.43
3	0.74	5.61	15	0.78	3.17
4	0.86	4.82	16	0.74	3.25
5	1.06	4.45	17	0.66	2.83
6	0.68	3.03	18	0.70	2.30
7	0.87	7.34	19	0.73	1.63
8	1.45	3.85	20	4.98	15.02
9	0.77	7.43	21	0.83	2.97
10	0.53	1.70	22	0.75	3.29
11	1.05	1.52	23	0.77	2.27

This decrease by a factor of four between indirect χ^2_{STAT} and χ^2_{GAF} demonstrates the necessity of using a dynamic description such as the SF-GAF to correctly reproduce experimental data. The value of the reduced indirect χ^2 gives important support to the weighting procedure developed in this method.

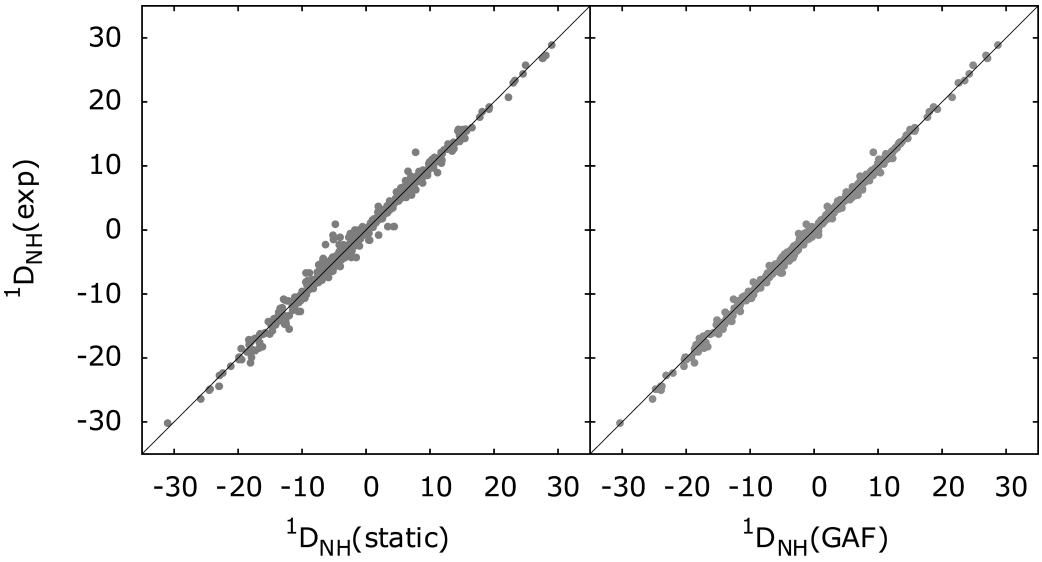

Figure 22 – Cross validation of the SF-GAF analysis of local motions in Ubiquitin, with randomly removed pairs of $^1D_{NH}$ coupling. Passive data reproduction assuming a static description (left) or using the SF-GAF analysis (right).

Statistical tests were made on direct χ^2 analysis. Direct χ^2 using static, S and SF-GAF model are summarized in Table 5, with the corresponding AIC_c. The improvement from a static to a S description and then to a SF-GAF description is highly significant: the improvement in AIC_c corresponds to have the SF-GAF description more than one million times more likely than the S. A similar conclusion can be obtained through F-tests.

Table 5 – Ubiquitin static, S and SF-GAF statistical analysis.

Model	χ^2	AIC_c
Static	4364.62	2026.00
S	1967.01	799.30
SF-GAF	1316.00	244.60

4.3.7 *Structural Information Content*

In addition to parameterizing the dynamic behaviour of each peptide, the mean orientations of the each in-plane internuclear vector can be extracted from the SF-GAF analysis. This has been done for N_i-H_i^N, C'_{i-1}-H_i^N, C'_{i-1}-N_i and C^{α}_{i-1}-C'_{i-1} vectors, and results are compared to the mean orientations extracted from the high resolution NMR structure 1d3z of Ubiquitin [175]. In order to simply overcome the intrinsic degeneracy of the RDCs, all vectors were folded into a sixteenth of the orientational space. Results are shown in

Figure 23. The correlation between the orientations underlines the ability of the SF-GAF approach to accurately determine peptide plane orientation. Of course, one can expect differences between the two sets of orientations. The 1d3z structure is a static description of the data and therefore the structure already reproduces all experimental data. The orientations obtained through SF-GAF correspond to the mean orientation of a continuous distribution of orientations that aim to represent conformational flexibility of the system at timescales up to the millisecond. We note that the Monte-Carlo analysis applied here provides a noise-based assessment of the orientational precision of the dynamically averaged mean conformation, allowing the determination of dynamically averaged orientations with their associated uncertainty.

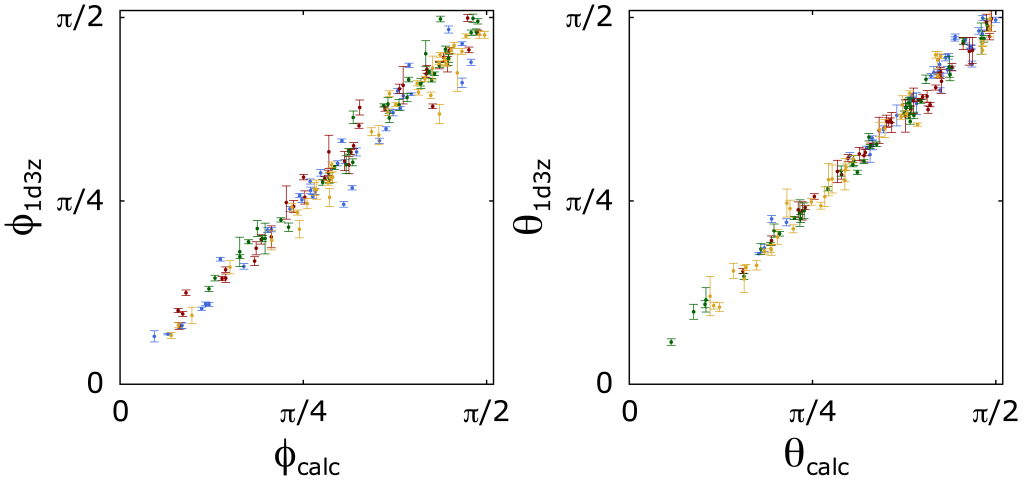

Figure 23 – Comparison between the mean orientation of in plane vectors extracted from SF-GAF analysis and the orientation determined from 1d3z Ubiquitin high resolution structure. Color coding correspond to the following vectors: N_i-H_i^N (blue), C'_{i-1}-H_i^N (red), C'_{i-1}-N_i (green) and C^{α}_{i-1}-C'_{i-1} (yellow).

4.4 CONCLUSION

Different conclusions can be drawn from the analysis, concerning the methods and the results. First for the methods:

- This approach allows the characterization of all of the alignment tensor properties including the amplitude. By comparing to simulated data, this characterization was shown to be quantitative (on the order of 1%) and absolute in the sense that no reference to any other experimental techniques such as ^{15}N relaxation is needed (this comparison of course assumes that the MD simulation used to test the approach has some validity).

- Following this analysis of the tensors, an accurate and quantitative characterization of local dynamics should be possible. Results were

coherent with ^{15}N relaxation, molecular dynamics and SCRM analysis, and extensively cross-validate, all of these elements giving cumulative support to the model.

- In an entirely noise free system, the only dynamic modes that would be missed by this approach are equal amplitude isotropic contributions common to the motion of all internuclear vectors. This may account for the 1% difference between known and estimated tensors observed in the bench mark simulations. In the real case this contribution cannot be known, as it is invisible to all — at least present — RDC-based approaches. MD simulations, and to an extent common sense, indicate that if present this motion does not represent a major contribution to the dynamics (maybe 1 or 2%). In the absence of experimental evidence for such motion, no justification was found for artificially introducing such a component.

On the `SF-GAF` results and their comparison to other techniques:

- Ubiquitin does not seem to present significant pervasive motion on timescales longer than the global rotational correlation time of the protein (~ 5 ns). Rather slower motions are heterogeneously present, and mainly situated in non-secondary structural elements on the surface of the molecule.

- In particular the N-terminal β-hairpin (residues 8 to 12) shows slower dynamics. Motion is also sampled by the MD simulation, which suggests that this motion mainly occurs within the tens of nanosecond time window. Nevertheless using three techniques sensitive to different timescales, more dynamics is systematically found in this region, as a function of increasing time-window. This suggests the presence of a motion occurring on a large continuous range of timescales, rather than, for example, the combination of a fast local motion occurring in different slowly exchanging local minima.

- The motion in the region of the turn 62-65 is more intriguing. Neither ^{15}N relaxation nor MD simulations are able to reveal enhanced dynamics in this regions, whereas the two `SF-GAF` and SCRM analyses clearly detect this. The distribution of α- and γ-motions in this region contrasts with the rest of the molecule, maybe indicating a collective motion that requires more sophisticated models than the one presented here. This region contains a very thermolabile hydrogen bond, which leads to one of the weakest detected $^{3}J_{C'N}$ trans-hydrogen bond scalar couplings [177] (see Figure 19). Therefore, two possible types of explanation can be invoked, which are not exclusive. Firstly, MD

simulations were done at 25 °C, which correspond to the temperature of ^{15}N relaxation measurements. RDCs measurements were mainly realized at 35 °C. Thus a motion can occur at high temperature that is quenched in the presence of the weak hydrogen bond. Secondly, the motion may occur at timescales slower than the timescales probed by the MD simulations.

Further improvement to our understanding of these slower motions is of course possible. The molecular dynamics simulation presented here does not cover the complete timescale probed by the RDCs, and further enhanced sampling may improve agreement. Investigation of other systems using this approach should also give further insights into the general nature of protein dynamics at slow timescales. Finally, one of the strengths of the GAF model lies in its ability to describe all dynamic information in terms of local and independent amplitudes of motion for each peptide plane. Nevertheless, this leads to the possible loss of information concerning the collective nature of the motion. The possibility to characterize this kind of motion should also be investigated.

5

ACCELERATED MOLECULAR DYNAMICS STUDY OF UBIQUITIN

ABSTRACT

Classical molecular dynamics simulations are currently unable to probe slow timescale motions in proteins. In order to overcome this limitation a method is developed to artificially enhance the conformational sampling by biasing the conformational energy potential. This bias induces a loss of the timescale information, which is retrieved by comparing the ensembles obtained through this restraint-free approach to experimental RDCs and J-couplings. Using this method a conformational ensemble representing Ubiquitin dynamics occurring on timescales up to the millisecond is obtained. The results converge very well with those of the SF-GAF analysis.

5.1 INTRODUCTION

As presented in Chapter 3 , MD simulations that can statistically sample the conformational landscape explored by a protein at timescales up to the millisecond are not currently available. One of the promising ways to explore such potentially complex landscapes resides in the so-called Accelerated Molecular Dynamics (AMD) [178]. This approach pioneered by Hamelberg et al., has been already successfully applied for various system [179–181]. The philosophy of this approach is to bias the energy landscape in order to increase the rates of interconversion between local energy minima. It is thus natural to try this approach to give a complementary view to the SF-GAF RDCs interpretation.

The aim of this chapter is not to further develop AMD methodology but rather to use it as a tool to generate conformational ensemble representative of Ubiquitin dynamics present in the RDCs probed time window. Adaptation of AMD protocol to Ubiquitin and generation of all the trajectories presented in this chapter, from which ensembles of structures are extracted,

were performed by PHINEUS RL MARKWICK. Even if the methods is presented in order to give the necessary basis for further interpretation, the emphasis will be set on the interpretation of these results in the context of RDC based studies and especially the SF-GAF presented in the Chapter 4.

5.2 PRINCIPLE AND METHODS

5.2.1 *Accelerated Molecular Dynamics Principle*

The AMD approach accelerates the exploration of the conformational landscape [178, 181]. In order to increase the rates of transition between two low energy conformational sub-states, a continuous non-negative bias potential is added to the potential surface energy. This modification is achieved according to two parameters: E_b a the so-called boost energy and α the acceleration parameter. The boost energy represent the upper limit for biasing energy landscape: the part of the energy landscape above this value will remain unchanged whereas all the area under this limit will be biased by adding a supplementary term to the potential. The obtained potential energy, for any position $\mathbf{r}$, in the conformational landscape is thus defined as:

$$V_{AMD}(\mathbf{r}) = \begin{cases} V(\mathbf{r}) & \text{if } V(\mathbf{r}) \geqslant E_b \\ V(\mathbf{r}) + \Delta V(\mathbf{r}) & \text{otherwise} \end{cases} \tag{5.1}$$

using:

$$\Delta V(\mathbf{r}) = \frac{\left(E_b - V(\mathbf{r})\right)^2}{\alpha + E_b - V(\mathbf{r})} \tag{5.2}$$

The result of such a bias is illustrated in Figure 24:

One major issue is to properly adjust the level of acceleration, i.e. calibrating E_b and α. Conformational sampling is enhanced by increasing E_b or decreasing α. Clearly this will decrease energy barrier heights as soon as the energy minimum of the considered transition remains below the boost energy. Oversampling resulting in a too flat energy landscape has to be avoided: this undesired situation will lead to a random walk through phase space, global decrease in the order parameters and possibly unfolding. One of the interesting characteristics of AMD is that thermodynamic and various equilibrium properties of the system can be accurately determined yielding a canonical average of an observable. The canonical ensemble is a statistical ensemble [107, 182] where the system is able to exchange energy with a thermal bath (the surroundings).

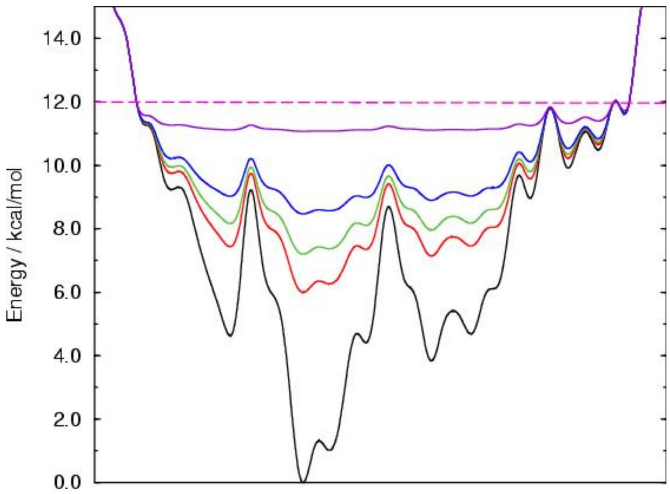

Figure 24 – AMD biased one dimensional energy landscape. This schematic representation show the effect of modifying the acceleration parameter α value. Boost energy is indicated with the horizontal dashed line.

The canonical variables of such an ensemble are N the number of particles, V the volume of the system and T the temperature. This weighing of any particular configuration in a canonical ensemble, e.g. the probability of a micro-state i is given by:

$$p_i = \frac{1}{\mathcal{Z}} e^{-E_i/k_B T} \tag{5.3}$$

with E_i the energy of the micro-state, k_B the Boltzmann constant and $\mathcal{Z}$ the partition function of the system defined as:

$$\mathcal{Z} = \sum_{i=1}^{N} e^{-E_i/k_B T} \tag{5.4}$$

Correcting the AMD canonical ensemble average to retrieve the unbiased one is obtained by re-weighting each micro-state with a $exp(V(\mathbf{r})/k_B T)$ factor.

This statistical weighting is then used to cluster solutions and the obtained low energy conformational sub-states are used as a starting point for standard short molecular dynamics simulations. This final ensemble of structures does not present local distortion due to modified potential but does present the conformationally enhanced properties derived from the biased potential acceleration.

5.2.2 *Simulation Details*

Simulation temperature was set to 300 K and the system was placed in a periodically repeating box with 6500 water molecules with a Langevin thermostat [142] and a Berendsen weak-coupling pressure-stat [183]. Electrostatic interactions were treated using the Particle Mesh Ewald [142, 184]. The force field used is AMBER ff99SB. All simulations were performed using the AMBER8 code [185].

The simulation is started from X-ray crystal structure of Ubiquitin (PDB code 1UBQ). After equilibration, ten different 5 ns MD simulations were started. Resulting structures are used as a starting point for the potential-biased simulation. Those trajectories are used as control trajectories and referred as null boost energy case.

Accelerated molecular dynamics were run at 8 levels of acceleration. The acceleration parameter were fixed at the standard value of 60 kcal·mol^{-1} [181] and the boost energy was set at 100, 150, 200, 250, 300, 350, 400 and 450 kcal·mol^{-1} above the dihedral angle energy (estimated from the average dihedral angle energy from the unbiased 5ns MD simulations). For each level of acceleration the complete procedure was repeated 20 times.

After re-weighting to the correct canonical Boltzmann distribution, a clustering analysis was performed on each AMD trajectory, using principal components analysis. A series of short 3 ns classical MD simulations were seeded from the resulting cluster (the initial 0.5 ns were discarded). From those trajectories a large free energy weighted structural ensemble, was extracted. This ensemble, which contains 20 450 structures, will be referred in the following as the AMD ensemble.

5.2.3 *Selection of the Level of Acceleration*

Accelerating conformation sampling is a useful tool for studying slow-dynamics via sensitive measurements, e.g. RDCs, but a crucial issue remains the estimation of the timescales probed with such an approach. In fact decreasing energy barrier height via AMD results in a non-trivial modification of transition rates. Instead of trying to characterize this effect from potential modification, the approach proposed here relies on experimental data to determine the optimal level of acceleration. The idea is to probe motions at timescales up to the millisecond and therefore comparison to RDCs, J-couplings and chemical shifts, all sensitive to this time window, can be made. Given a set of applied accelerations, the resulting ensembles will not similarly reproduce experimental data. The one that best fits a

given set of homogeneously time sensitive data should present the best agreement in terms of probed timescales. In fact, if the only difference between different AMD ensembles is the level of acceleration the ensembles should all represent the dynamics of the same system but sampled at times windows increasing with the level of acceleration.

As alignment media were not explicitly incorporated in the simulation, AMD generated-ensembles are directly compared to the experimental data using an SVD based approach in order to extract the optimal tensor for each molecular ensemble. SVD principle is presented in Annexe A. A similar approach was used when treating J-couplings. A Karplus equation to parametrize those couplings as a function of backbone ϕ-angle can be used:

$$^3J_{AB} = A_K \cos^2(\phi + \theta_{AB}) + B_K \cos(\phi + \theta_{AB}) + C_K \tag{5.5}$$

with A_K, B_K and C_K the Karplus parameters and θ_{AB} an offset angle which is typically around 180° for $^3J_{H^N C'}$, −60° for $^3J_{H^N H^\alpha}$ and 60° for $^3J_{H^N C^\beta}$. Using a matricial form of this equation, for a set of N for ^{3}J-couplings, the AMD ensemble best fitting Karplus coefficient can be obtain though SVD of the following system:

$$\begin{pmatrix} \left\langle \cos^2(\phi_1 + \theta_{AB}) \right\rangle & \left\langle \cos(\phi_1 + \theta_{AB}) \right\rangle & 1 \\ \left\langle \cos^2(\phi_2 + \theta_{AB}) \right\rangle & \left\langle \cos(\phi_2 + \theta_{AB}) \right\rangle & 1 \\ \vdots & \vdots & \vdots \\ \left\langle \cos^2(\phi_N + \theta_{AB}) \right\rangle & \left\langle \cos(\phi_N + \theta_{AB}) \right\rangle & 1 \end{pmatrix} \begin{pmatrix} A_K \\ B_K \\ C_K \end{pmatrix} = \begin{pmatrix} {}^3J_{AB,1} \\ {}^3J_{AB,2} \\ \vdots \\ {}^3J_{AB,N} \end{pmatrix} \tag{5.6}$$

The θ_{AB} are then optimized by repeating the analysis by changing θ_{AB} by 1° steps.

The dataset used for the level of acceleration selection comprises 23 $^1D_{NH}$ couplings datasets. They all come from the literature (see Section 4.2.1) only RDC datasets with more than 40 $^1D_{NH}$ were used and the two alignment media used to optimize 1d3z structure [68] were not used (in the interest of comparison to a static model). All analyses excluded residue 5 for which particularly poor results where constantly found and resulting S^2_{NH} much higher than the one extracted from ^{15}N relaxation. This unexpected behavior was assumed to be a force-field issue.

5.2.4 Q_f *and* R_f *Factors*

Q_f and R_f factors are two similar measures of the quality of data reproduction. Contrarily to χ^2 estimation they are defined to be independent of experimental error, as the data reproduction is estimated through:

$$Q_f = \sqrt{\frac{\sum_i (X_i^{calc} - X_i^{exp})^2}{\sum_i (X_i^{exp})^2}} = \sqrt{2}\, R_f \tag{5.7}$$

where i runs over all considered experimental data.

5.3 RESULTS AND DISCUSSION

5.3.1 *Data Reproduction and Level of Acceleration*

The S^2_{NH} obtained from the AMD ensemble corresponding to different levels of acceleration and the corresponding data reproduction can be found in Figure 25.

The most appropriate level of acceleration is obtained for a boost energy of $250\,\mathrm{kcal{\cdot}mol^{-1}}$, at which the data reproduction of both RDCs and J-couplings is at its best. Using less acceleration samples too narrow conformational space, whereas further increasing the boost energy will spread obtained structures over a too broad conformation space.

At this level of acceleration the RDCs cumulative R_f, $R_{cum}(RDC)$ over all the 23 alignment media is 2.496 which correspond to an averaged $R_f(RDC)$ $\langle R_f(RDC)\rangle = 0.109$ and varying quite homogeneously from media to media between 0.090 and 0.129. Typical data reproduction is illustrated in Figure 26.

Comparison of data reproduction using the AMD ensemble compared to that obtained through a control MD is presented in Figure 27. Except for residue 54 the data reproduction is improved in the AMD ensemble. This improvement spreads all along the sequence: in both dynamic regions, such as loop 8-12 (see Section 4.3) or more structured ones e.g. residues 15, 17, 34, 42 and 67. For the former this improvement mainly came from the more appropriate conformational sampling, whereas for the latter it arises from the more adapted tensor representation with this improved ensemble-average.

At the same optimal level of acceleration, J-couplings data reproduction is also improved, although compared to $^1D_{NH}$ couplings the J-coupling

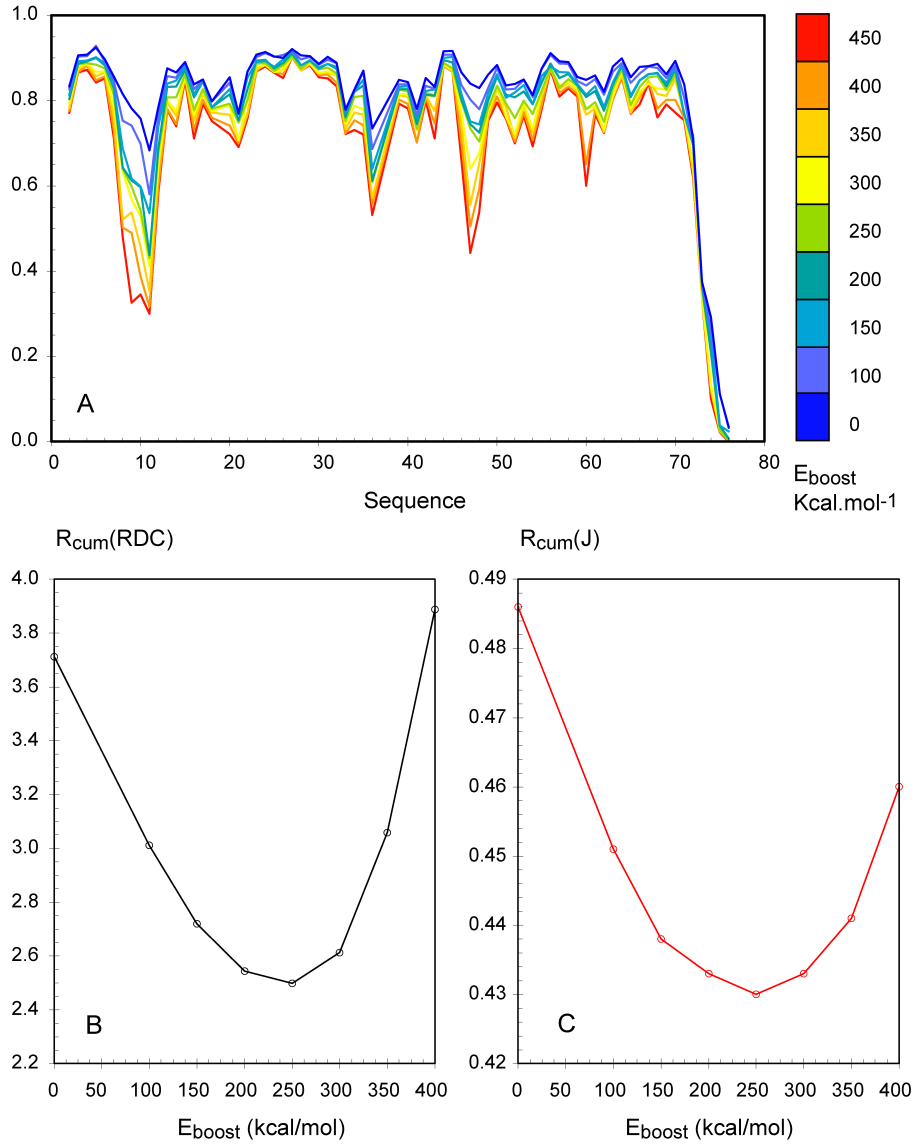

Figure 25 – Effect of increasing the acceleration level in Ubiquitin AMD. From top to bottom, the boost energy is set at 0 (standard 5 ns MD control set), 100, 150, 200, 250, 300, 350, 400, and 450 kcal·mol^{-1}. The acceleration parameter, α, was fixed at a value of 60 kcal·mol^{-1}. (A) S^2_{NH} order parameters obtained for increasing boost energy. Change in the trajectory-averaged cumulative R_f value for RDCs $R_{cum}(RDC)$ (B) and J-couplings $R_{cum}(J)$ (C) as a function of the acceleration level.

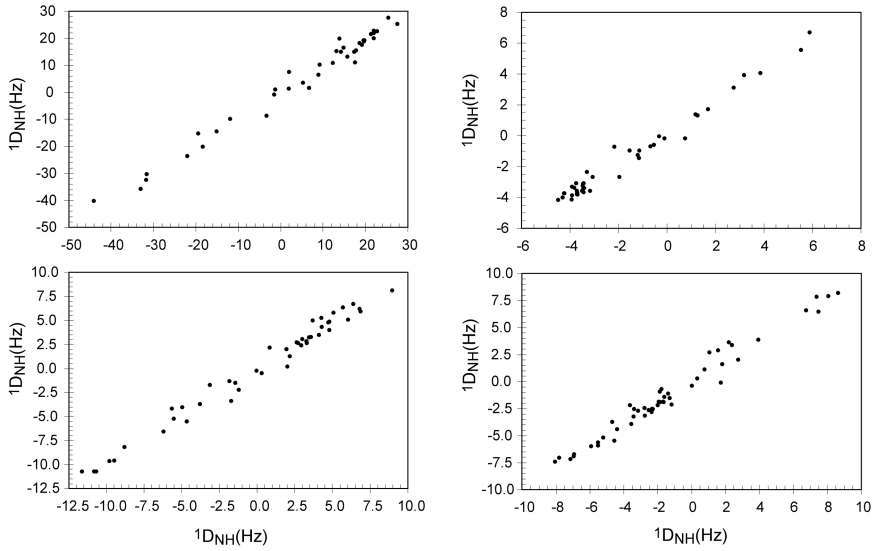

Figure 26 – Typical RDCs data reproduction using AMD ensemble. Experimental vs AMD calculated $^{1}D_{NH}$ couplings for four representative alignment media . The trajectory averaged $R_f(RDC)$ for the shown alignment media are respectively 0.096, 0.098, 0.100, and 0.111.

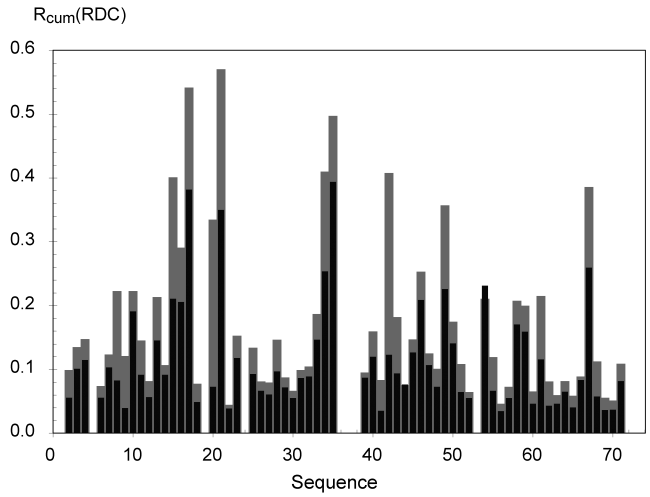

Figure 27 – Residue specific trajectory averaged RDC cumulative R-factors $R_{cum}(RDC)$ obtained using a standard MD approach (black), compared to the AMD approach (grey)

data reproduction appears to be less sensitive to the level of dynamics (see Figure 25). Data reproduction and extracted Karplus parameters for $^3J_{H^N H^\alpha}$ are shown in Figure 28. These Karplus parameters are compared to those extracted from a single structure (1d3z) analysis, from a standard 5 ns MD and from sum-overstates density functional theory (SOS-DFT) calculated for model peptides. There is a clear tendency to converge towards the DFT curve with increasing dynamics: the AMD curve being almost identical to the DFT one. This is in agreement with previously described tendency where fitting Karplus parameters absorbed part of the motion.

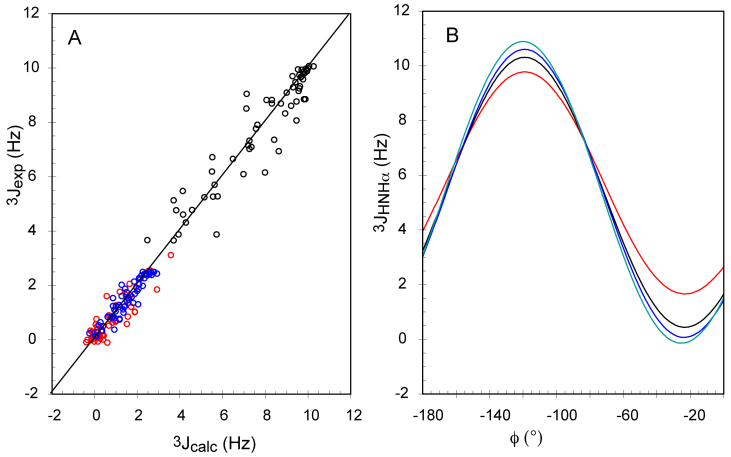

Figure 28 – J-coupling data reproduction using AMD ensemble. A: Experimental vs AMD calculated scalar J-couplings: $^3J_{H^N H^\alpha}$ (black circles), $^3J_{H^N C^\beta}$ (red circles), and $^3J_{H^N C'}$ (blue circles). (B) $H^N H^\alpha$ Karplus Curves. Red: optimal Karplus curve for a static description (1d3z). Black: optimal Karplus curve for standard 5ns MD simulation. Blue: optimal Karplus curve for optimal AMD result. Cyan: DFT Karplus curve for model peptide.

5.3.2 *Order Parameters*

Even if the AMD ensemble represent dynamics at timescales up to the millisecond it is possible to extract from it order parameter corresponding to faster timescales. In fact the acceleration procedure results in a large ensemble of structures distributed in a set of energy minima. All this sub-ensemble correspond to conformational region that are in exchange on timescales that are fast on the chemical shift timescale and slower than the molecular correlation time. Each of those minima can contribute to fast relaxation processes with a weight fixed by this relative probability.

Therefore a description of fast dynamics (ps-ns) was obtained by averaging order parameters extracted from each of the separate energy substates. Order parameters extracted from AMD to represent fast timescales

or timescales up to the millisecond are shown in Figure 29. As previously described for protein GB3 [181], the obtained fast timescales S^2_{NH} order parameters are in good agreement with ^{15}N relaxation data, and better than results obtain using a standard MD approach (data not shown). Nevertheless a small bias towards too high order parameters can be observed. Concerning slow dynamics the S^2_{NH} profile gives a good convergence between $S^2_{NH,REL}$ and $S^2_{NH,AMD}$, suggesting the absence of slow dynamics in secondary structures. A small excess of dynamics is found in the N terminal β-hairpin (residues 8 to 12), which indicates the presence of additional dynamics at slower timescales.

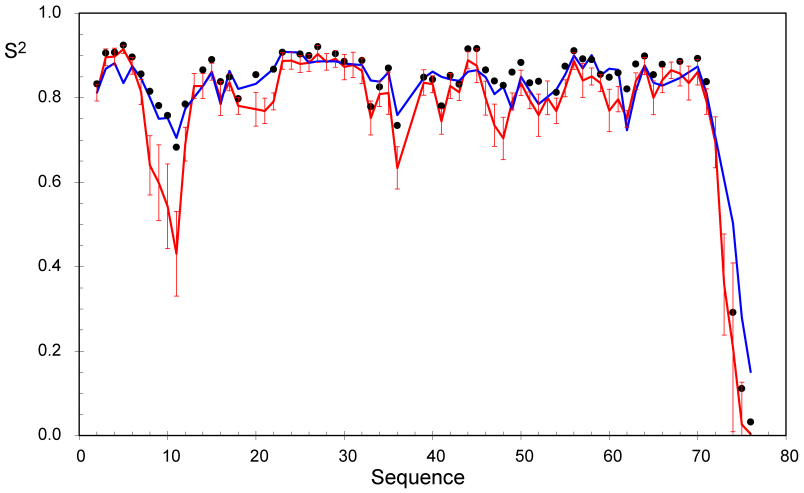

Figure 29 – N_i-H^N_i Order parameter S^2_{NH} corresponding to fast timescales or timescales up to the millisecond according AMD ensemble: S^2_{NH} from ^{15}N relaxation experimental data (blue line), fast time-scale (ps-ns) S^2_{NH} from AMD (black circles) and the slow time-scale S^2_{NH} (red line). The error bars correspond to the standard deviation of the S^2_{NH} extracted from the 20 AMD trajectories.

5.3.3 *Conformationally Sampled Space and Comparison with Others Approaches*

Comparison of the AMD ensemble to static structures (NNR 1d3z [175] and X-rays 1ubq [186]) indicates that the AMD ensemble results in an average structure closer to the 1d3z static model and that the dispersion of the conformational sampling about this mean increases compared to a standard MD. A representative ensemble of structures taken from the AMD ensemble is shown in Figure 30.

More interestingly AMD description can be compared to other NMR RDC-derived structure ensembles, such as the EROS ensemble [139]. This ensemble, which consist of 116 structures obtained with an averaged restrained molecular dynamics approach, was restrained using extensive nOe restraints

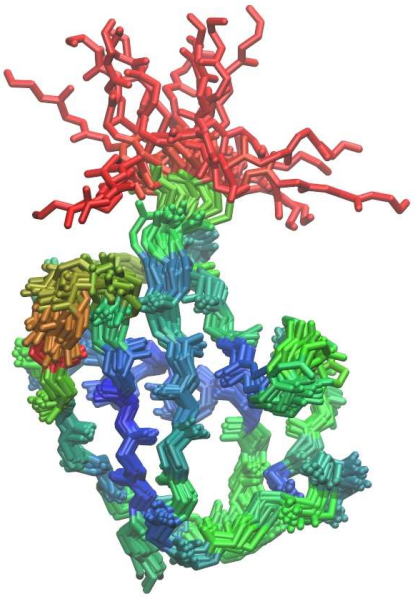

Figure 30 – Twenty-four representative structures taken from the AMD ensemble. Residues are color-coded according to the $^{1}D_{NH}$ order parameters $S^2_{NH,AMD}$ from $S^2_{NH,AMD} = 1.0$ (blue) to $S^2_{NH,AMD} = 0.0$ (red).

and $^{1}D_{NH}$ RDCs. Principle component analysis of the two ensembles exhibits similar sampling of the conformational landscape even though the comparison is made more complex by the quite different number of structures used, and by the fact that the EROS ensemble is not free energy weighted.

Data reproduction between the different approaches can be checked but difficulties arise from the fact that the AMD approach is restraint free and that both static or ensemble averaged actively use experimental NMR data (except the X-ray structure). Without entering into details the AMD approach gives better data reproduction that 1ubq, similar to 1d3z and a bit worse than EROS, a very promising result as the active use of a given dataset will obviously improve drastically its reproduction.

Concerning order parameters, S^2_{NH}, AMD and EROS ensemble can be compared. In the case of EROS ensemble, the authors estimate that the use of 116 structures was not enough to properly describe libration of N_i-H^N_i internuclear vector and therefore applied an overall scaling factor of 0.93 in order to correct this defect. The two approaches nevertheless present good agreement in terms of the relative S^2_{NH} order parameter distribution, and prior to applying this scaling, are quantitatively similar.

5.4 COMPARISON WITH STRUCTURE-FREE GAF ANALYSIS

5.4.1 *Complementarity of the two approaches*

Before starting to compare the to approach presented here, the AMD and the SF-GAF (see chapter 4), it is maybe better to reexamine the fundamentals of the two approaches.

The SF-GAF is based on a biophysical model that allows an analytical expression of dynamically averaged RDCs. Extraction of the dynamic parameters is achieved by fitting adjustable parameters to the experiment available RDCs. Only RDCs are used in this study, providing coherent time window sensitivity. No force-field is used, even if one may see the fixed topology of the peptide plane as a force-field, as the geometry is absolutely unmodifiable it should be more considered as a hypothesis of the model. All residues are treated independently, avoiding error propagation between planes but allowing any level of fluctuation from one site to its neighbour.

The AMD approach relies on a purely molecular dynamics based protocol. Therefore it is sensitive to force field defects but the presence of essentially well calibrated potential terms will hopefully ensure the coherence of the obtained ensemble. In fact force fields are parametrized to reproduce experimental data and all the knowledge coming from this parameterization is implicitly present in the MD protocols. Moreover Ubiquitin is treated as an entire molecule and therefore neighbour-dependent interactions will be completely different to that present in SF-GAF: generally smoother evolution of parameters should be expected in the AMD. Is worth emphasizing that the AMD approach is restraint-free and experimental data are only used to select the correct level of acceleration, that is to say the timescales probed by the ensembles.

5.4.2 *Order Parameters*

Order parameters for N_i-H_i^N, C'_{i-1}-N_i and C^{α}_{i-1}-C'_{i-1} vectors (S^2_{NH}, S^2_{CN} and S^2_{CC}) are presented in Figure 31. Concerning N_i-H_i^N order parameters the agreement in both shape and amplitude is very good, leading to almost indistinguishable values within the experimental error. Concerning loop region 8-12 which exhibits the most slow dynamics order parameters are almost identical. In secondary structured regions an alternation of a slightly more and slightly less dynamics is observed for one approach compared to the other with similar averaged values. A bit more dynamics is obtained through SF-GAF analysis in residues 51-53. The only major discrepancy is present in the 62-65 loop region.

Concerning C^{α}_{i-1}-C'_{i-1} vectors which are almost perfectly orthogonal to the N_i-H^N_i vector differences are bigger. In the N-terminal β-hairpin the residue 11 is found to be more dynamics in `SF-GAF` analysis. Slower $S^2_{CC,GAF}$ can be detected for residues 5, 25, 35-36, 42, 52 and 64-65. For all those residues, except 35, `SF-GAF` selected model was `M-III` (all the three reorientation diffusion angles are optimized). Otherwise very good agreement appears.

For C'_{i-1}-N_i vector the situation is intermediate between the two previously described N_i-H^N_i and C^{α}_{i-1}-C'_{i-1} vectors, which is normal as the C'_{i-1}-N_i vector can be expressed as a linear combination of the two nearly orthogonal N_i-H^N_i and C^{α}_{i-1}-C'_{i-1} vectors.

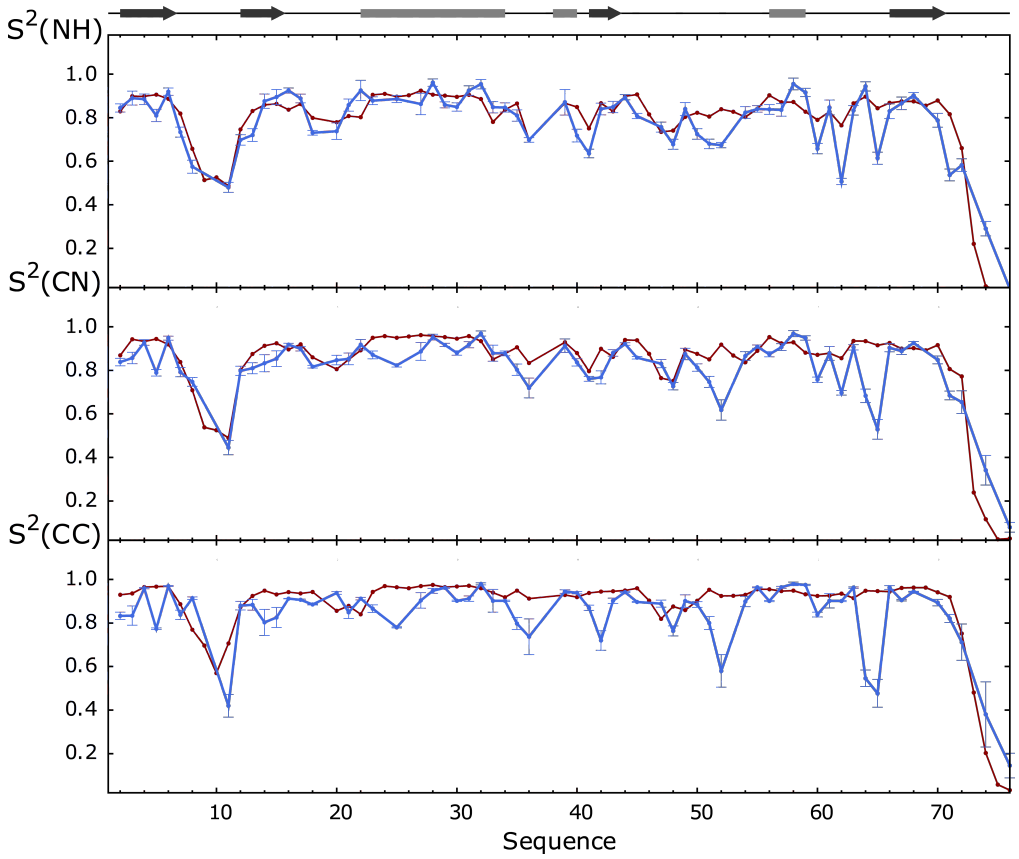

Figure 31 – Order parameters comparison of the `SF-GAF` and the AMD approaches. From top to bottom: N_i-H^N_i, C'_{i-1}-N_i and C^{α}_{i-1}-C'_{i-1} order parameters. `SF-GAF` results are shown in blue, AMD ones in red. Grey boxes indicate α-helix and darker arrows indicate β-sheet.

First of all considering the differences between the two methods the agreement is remarkably good, giving further support to the two approaches. Nevertheless some differences are revealed. In general better convergence is found for the two approaches for S^2_{NH} than for S^2_{CC}. This seems to be completely coherent with the following points:

- Firstly AMD and especially SF-GAF have poorer precision to determine those order parameters S^2_{CC} as this direction of the peptide plane exhibits less dynamics than the N_i-H^N_i vector, the C^α_{i-1}-C'_{i-1} vector being completely insensitive to γ-motion.

- Most importantly the used dataset contains an overwhelming majority of $^1D_{NH}$ couplings. The amounts of RDCs that exactly probed this direction is tiny as only two alignment media contained $^1D_{C'C^\alpha}$ and even $^1D_{C'H^N}$ and $^1D_{C'N}$ that are partially informative. Moreover $^1D_{C'C^\alpha}$, $^1D_{C'H^N}$ and $^1D_{C'N}$ present an experimentally smaller range and therefore experimental precision is often reduced.

Resulting effects can be clearly seen in the accuracy estimation through Monte-Carlo calculations for SF-GAF analysis, where S^2_{CC} order parameters are definitely less accurately defined than the S^2_{NH} ones.

Nevertheless, extended motion in the SF-GAF model can have a physical meaning too. For example residue 5 recurrently exhibits too small dynamics in the AMD approach, which is not observed in SF-GAF model. Concerning C-terminal loop 62-65 the difference is present for all in plane vectors. As discussed in Section 4.4 different reasons can explain this discrepancy. In the light of AMD calculations this seems not to be due to slow motion occurring 25 °C. Considering the fact that fast dynamics does not exhibit differences in this loop over a range of temperature up to 40 °C [176] and that AMD is not able to detect this motion this discrepancy can be explained, if not due to a defect of one approach, by a slow motion occurring at higher temperature (above the range of stability of the thermolabile hydrogen-bound [177]).

5.5 CONCLUSION

In this chapter the Accelerated Molecular Dynamics was presented as a tool to generate ensemble representative of timescales up to the millisecond. This is a completely restraint free approach that enhances conformational sampling by biasing the potential in order to get easier transition over high energy transition barriers. The bias induces the loss of timescales information as the modification of the potential perturbs any transition rate. This lost time information can be reintroduced by estimating the level of acceleration that best reproduces a experimental homogeneously time sensitive set of data. Here the use of RDCs and J-couplings leads to a ensemble representative of timescales up to the millisecond.

This ensemble, considering data reproduction and by comparison with other approaches has been shown to give realistic description of dynamical

behavior of Ubiquitin. This is especially important in comparison to the SF-GAF approach. Due to their intrinsic differences any convergence of the two approaches can be seen as a good support for both methods. In fact the SF-GAF, an analytical model fitted to experimental data and the AMD ensemble generated by a purely computational method, are two diametrically opposed ways to characterize Ubiquitin dynamics. Even if some small rationalizable differences remain the convergence of the two methods, considering for example order parameters is clear. The similar site specific distribution confirms the accuracy of both approaches to site specifically estimate dynamics and the convergence to the same absolute level of dynamics gives a robust determination of the amount of conformational flexibility present in folded proteins.

6

PROTEIN GB3 DYNAMIC ANALYSIS: TOWARDS A DESCRIPTION OF COMMON MOTIONS

ABSTRACT

NMR measurable parameters are often site specific. Thus, models of dynamics, including GAF models, generally treat motions using an entirely local description. In this study, new analytical models are developed and applied to the identification of collective motional modes in protein GB3. In a first step the SF-GAF method is used to determine the amount of dynamics in the system, then more complex models of motion are used to investigate the presence of common motions. The possibility of using a description that explicitly discriminates between collective and local motions appears to be possible in the β-sheet of protein GB3 despite the simplicity of the model used for the common motion. This example can be seen as a proof of principle and these descriptions are expected to be possibly more relevant for investigating systems where complex dynamic occurs.

6.1 INTRODUCTION

As presented in Chapter 4, GAF methods can be used to accurately characterize local motion in folded proteins. This description is applied plane by plane, in a site specific way, with all the dynamics occurring at a given peptide plane being interpreted in terms of three local amplitudes of reorientation.

This interpretation is a useful tool for characterizing the level of dynamics present at each peptide unit and to determine the directionality of this motion. Nevertheless, the possible presence of motions shared by different planes is not explicitly taken into account. If both a collective and an individual motion occur, they will both be integrated into the local GAF description.

Biologically important slow dynamic modes may however be expected to be collective, requiring larger activation energies that may be responsible for the slower timescale of the motion. Collective motions are thought to be responsible for information transport and allosteric regulation, controlled by large-scale conformational changes and domain motions. Normal Mode Analyses [187–189] and Coupled Oscillators Models [190] derived from static structural models also support the physically intuitive presence of common components of motional modes. It would therefore appear to be important to be able to include a component of collective motional reorientation in the GAF analysis.

Indeed, a previously published analysis of protein GB3 (see Annexe B) using the GAF models already identified correlated motions within the β-sheet, on the basis of repetitive alternating γ- and β-motions and $^{3}J_{C'N}$ trans-hydrogen-bond scalar coupling data [164, 167]. Recent solid state NMR relaxation rate R_1 measurements in protein GB3 also suggest the presence of such a motion in the solid state [191]. The solution state observations therefore already revealed a key drawback of all site-specific interpretations of protein backbone dynamics, namely that local characterization of motion assuming independent motional modes — the simplest assumption — can only indicate the possible presence of collective modes, and in some cases may even mask their presence. In order to take this approach further, we have therefore investigated the use of the GAF model to characterize commonly shared motions in the β-sheet of this protein. In this chapter, three novel models will be described, the 3-0D-GAF, 3-1D-GAF and 3-3D-GAF models, which allow the explicit expression of individual and collective motions in the GAF formalism. They will be applied and tested using existing extensive RDC datasets. The model that combines local and shared dynamics will be called LS-GAF.

6.2 GAF MODELS FOR BOTH LOCAL AND SHARED MOTIONS

3-0D-GAF, 3-1D-GAF and 3-3D-GAF models are analytical extensions of the 3D-GAF model. Instead of using a set of three amplitudes that characterize the total amount of dynamics present in a plane, the two models 3-1D-GAF and 3-3D-GAF propose to explicitly integrate two types of motions: a global shared motion and a local individual motion.

6.2.1 *Model Description*

The philosophy of these LS-GAF models is to describe a collective motion using a 3D-GAF description and a local motion through a γ-1D-GAF for the

3-1D-GAF description or a 3D-GAF for the 3-3D-GAF description. 3-0D-GAF model assumes only the presence of a global motion. For a given structural motif of the system, e.g. a β-sheet, an α-helix or a loop, a common motion is assumed and described using an extension of the 3D-GAF model.

In those models, a set of three orthogonal axes $\mathcal{A}_\delta$, $\mathcal{A}_\epsilon$ and $\mathcal{A}_\zeta$ is defined. As they have to describe the directionality of the anisotropic collective motion, their orientations can not be defined *a priori* and they are therefore present as adjustable parameters of the model. Corresponding reorientational diffusion dynamics about three orthogonal axes are defined by σ_δ, σ_ϵ and σ_ζ, each of those parameters determining the GAF amplitude of reorientation about the considered axis. As in the standard 3D-GAF description, the three diffusive reorientations are considered as independent and described by a Gaussian distribution centered on the average position of the structural motif. The $\mathcal{R}_\delta$, $\mathcal{R}_\epsilon$ and $\mathcal{R}_\zeta$ are the frame formed by the three $\mathcal{A}_\delta$, $\mathcal{A}_\epsilon$ and $\mathcal{A}_\zeta$ axes, with respectively the z-axis along $\mathcal{A}_\delta$, $\mathcal{A}_\epsilon$ and $\mathcal{A}_\zeta$.

The local dynamics is then defined using a γ-1D-GAF or 3D-GAF model, defined exactly as previously (see Section 2.6.3).

Global and local reorientations are supposed statistically independent, and the obtained peptide plane orientation corresponds to position averaged through all 3, 4 or 6 GAF motions.

The physical interpretation of these collective GAF motions may depend on the considered system. Here, only common direction and shared amplitudes are imposed. Considering a two domain system and using each domain as structural motif for those LS-GAF can be simply interpreted in terms of domain reorientation. Here, the structural motif is, for example, a β-sheet and the shared GAF motions are expected to describe global fluctuations that can be induced by a collective motion within the motif. Importantly, using the β-sheet as a motif does not assume that it reorientates as a rigid object: the fact that two planes share the same motion does not impose their relative motion, as neither translational information nor information concerning possible correlations are available.

6.2.2 *Analytical Derivations*

The 3-0D-GAF, 3-1D-GAF and 3-3D-GAF RDCs averaging effect can be described using averaged rank two spherical harmonics. Following the principles presented in Section 2.6.3, the independence of all GAF reorientations

allows successively averaging over all motions. We therefore obtain, omitting the `1D-GAF` subscript for clarity:

$$\left\langle Y_{2,p}(\theta,\phi)\right\rangle_{\texttt{3-3D-GAF}} = \left\langle\left\langle\left\langle\left\langle\left\langle\left\langle Y_{2,p}(\theta,\phi)\right\rangle_{\alpha}\right\rangle_{\beta}\right\rangle_{\gamma}\right\rangle_{\delta}\right\rangle_{\epsilon}\right\rangle_{\zeta} \tag{6.1}$$

$$\left\langle Y_{2,p}(\theta,\phi)\right\rangle_{\texttt{3-1D-GAF}} = \left\langle\left\langle\left\langle\left\langle Y_{2,p}(\theta,\phi)\right\rangle_{\gamma}\right\rangle_{\delta}\right\rangle_{\epsilon}\right\rangle_{\zeta} \tag{6.2}$$

$$\left\langle Y_{2,p}(\theta,\phi)\right\rangle_{\texttt{3-0D-GAF}} = \left\langle\left\langle\left\langle Y_{2,p}(\theta,\phi)\right\rangle_{\delta}\right\rangle_{\epsilon}\right\rangle_{\zeta} \tag{6.3}$$

Considering the three global reorientation averaging modes first, according to ζ-, ϵ- and δ-motion in this order, we obtain:

$$\begin{aligned}\left\langle Y_{2,p}(\theta,\phi)\right\rangle_{\texttt{3-3D-GAF}} = \sum_{q=-2}^{2}\Bigg[& e^{-iq\alpha_\zeta}\, d^{(2)}_{q,p}(\beta_\zeta)\, e^{-ip\gamma_\zeta}\, e^{-\frac{1}{2}q^2\sigma_\zeta^2} \\ & \times \sum_{r=-2}^{2}\Bigg[d^{(2)}_{r,q}\left(\frac{\pi}{2}\right) e^{-iq\frac{\pi}{2}}\, e^{-\frac{1}{2}r^2\sigma_\epsilon^2} \\ & \quad \times \sum_{s=-2}^{2}\Bigg[d^{(2)}_{s,r}\left(\frac{\pi}{2}\right) e^{-ir\frac{\pi}{2}}\, e^{-\frac{1}{2}s^2\sigma_\delta^2} \\ & \quad\quad \times \left\langle\left\langle\left\langle Y_{2,s}(\theta_\delta,\phi_\delta)\right\rangle_{\alpha}\right\rangle_{\beta}\right\rangle_{\gamma}\Bigg]\Bigg]\Bigg]\end{aligned} \tag{6.4}$$

$$\begin{aligned}\left\langle Y_{2,p}(\theta,\phi)\right\rangle_{\texttt{3-1D-GAF}} = \sum_{q=-2}^{2}\Bigg[& e^{-iq\alpha_\zeta}\, d^{(2)}_{q,p}(\beta_\zeta)\, e^{-ip\gamma_\zeta}\, e^{-\frac{1}{2}q^2\sigma_\zeta^2} \\ & \times \sum_{r=-2}^{2}\Bigg[d^{(2)}_{r,q}\left(\frac{\pi}{2}\right) e^{-iq\frac{\pi}{2}}\, e^{-\frac{1}{2}r^2\sigma_\epsilon^2} \\ & \quad \times \sum_{s=-2}^{2}\Bigg[d^{(2)}_{s,r}\left(\frac{\pi}{2}\right) e^{-ir\frac{\pi}{2}}\, e^{-\frac{1}{2}s^2\sigma_\delta^2} \\ & \quad\quad \times \left\langle Y_{2,s}(\theta_\delta,\phi_\delta)\right\rangle_{\gamma}\Bigg]\Bigg]\Bigg]\end{aligned} \tag{6.5}$$

$$\left\langle Y_{2,p}(\theta,\phi)\right\rangle_{\texttt{3-1D-GAF}} = \sum_{q=-2}^{2}\Bigg[e^{-iq\alpha_\zeta}\, d^{(2)}_{q,p}(\beta_\zeta)\, e^{-ip\gamma_\zeta}\, e^{-\frac{1}{2}q^2\sigma_\zeta^2} \times \sum_{r=-2}^{2}\Bigg[d^{(2)}_{r,q}\left(\frac{\pi}{2}\right) e^{-iq\frac{\pi}{2}}\, e^{-\frac{1}{2}r^2\sigma_\epsilon^2} \times \sum_{s=-2}^{2}\Bigg[d^{(2)}_{s,r}\left(\frac{\pi}{2}\right) e^{-ir\frac{\pi}{2}}\, e^{-\frac{1}{2}s^2\sigma_\delta^2} \times Y_{2,s}(\theta^0_\delta,\phi^0_\delta)\Bigg]\Bigg]\Bigg] \tag{6.6}$$

where the orientation $(\theta_\delta,\phi_\delta)$ of the vector of interest is expressed in the $\mathcal{R}_\delta$ frame and $(\alpha_\zeta,\beta_\zeta,\gamma_\zeta)$ the three Euler angles defining the rotation from the PAS frame to the $\mathcal{R}_\zeta$. For `3-0D-GAF` motion no more dynamics is assumed and thus $(\theta^0_\delta,\phi^0_\delta)$ corresponds to the orientation of the averaged internuclear vector in the $\mathcal{R}_\delta$ frame. Introducing the $(\alpha_\gamma,\beta_\gamma,\gamma_\gamma)$ angles which describe the rotation from $\mathcal{R}_\zeta$ to $\mathcal{R}_\gamma$ the following expressions can be obtained, by averaging for the `3-3D-GAF` in this order γ-, β- and α-motions:

$$\left\langle Y_{2,p}(\theta,\phi)\right\rangle_{\texttt{3-3D-GAF}} = \sum_{q=-2}^{2}\Bigg[e^{-iq\alpha_\zeta}\, d^{(2)}_{q,p}(\beta_\zeta)\, e^{-ip\gamma_\zeta}\, e^{-\frac{1}{2}q^2\sigma_\zeta^2} \times \sum_{r=-2}^{2}\Bigg[d^{(2)}_{r,q}\left(\frac{\pi}{2}\right) e^{-iq\frac{\pi}{2}}\, e^{-\frac{1}{2}r^2\sigma_\epsilon^2} \times \sum_{s=-2}^{2}\Bigg[d^{(2)}_{s,r}\left(\frac{\pi}{2}\right) e^{-ir\frac{\pi}{2}}\, e^{-\frac{1}{2}s^2\sigma_\delta^2} \times \sum_{t=-2}^{2}\Bigg[e^{-iq\alpha_\gamma}\, d^{(2)}_{t,s}(\beta_\gamma)\, e^{-is\gamma_\gamma}\, e^{-\frac{1}{2}t^2\sigma_\gamma^2} \times \sum_{u=-2}^{2}\Bigg[d^{(2)}_{u,t}\left(\frac{\pi}{2}\right) e^{-it\frac{\pi}{2}}\, e^{-\frac{1}{2}u^2\sigma_\beta^2} \times \sum_{\nu=-2}^{2}\Bigg[d^{(2)}_{\nu,u}\left(\frac{\pi}{2}\right) e^{-iu\frac{\pi}{2}}\, e^{-\frac{1}{2}\nu^2\sigma_\alpha^2} \times Y_{2,\nu}(\theta_{\alpha,0},\phi_{\alpha,0})\Bigg]\Bigg]\Bigg]\Bigg]\Bigg]\Bigg] \tag{6.7}$$

and more simply for the 3-1D-GAF model with local γ-motion:

$$\begin{aligned}\left\langle Y_{2,p}(\theta,\phi)\right\rangle_{\texttt{3-1D-GAF}} = \\ \sum_{q=-2}^{2}\Bigg[e^{-iq\alpha_\zeta}\, d^{(2)}_{q,p}(\beta_\zeta)\, e^{-ip\gamma_\zeta}\, e^{-\frac{1}{2}q^2\sigma_\zeta^2} \\ \times\sum_{r=-2}^{2}\Bigg[d^{(2)}_{r,q}\left(\frac{\pi}{2}\right)\, e^{-iq\frac{\pi}{2}}\, e^{-\frac{1}{2}r^2\sigma_\epsilon^2} \\ \times\sum_{s=-2}^{2}\Bigg[d^{(2)}_{s,r}\left(\frac{\pi}{2}\right)\, e^{-ir\frac{\pi}{2}}\, e^{-\frac{1}{2}s^2\sigma_\delta^2} \\ \times\sum_{t=-2}^{2}\Bigg[e^{-iq\alpha_\gamma}\, d^{(2)}_{t,s}(\beta_\gamma)\, e^{-is\gamma_\gamma}\, e^{-\frac{1}{2}t^2\sigma_\gamma^2} \\ \times Y_{2,t}(\theta_{\gamma,0},\phi_{\gamma,0})\Bigg]\Bigg]\Bigg]\Bigg]\end{aligned} \tag{6.8}$$

The expression of an in-plane RDC can be eventually obtained according to:

$$\begin{aligned}D^{\texttt{3-3D-GAF}}_{IS} = d_{IS}\sqrt{\frac{16\pi}{5}}\Bigg[A_a\left\langle Y_{2,0}\right\rangle_{\texttt{3-3D-GAF}} \\ +\sqrt{\frac{3}{8}}A_r\left(\left\langle Y_{2,-2}\right\rangle_{\texttt{3-3D-GAF}}+\left\langle Y_{2,2}\right\rangle_{\texttt{3-3D-GAF}}\right)\Bigg]\end{aligned} \tag{6.9}$$

$$\begin{aligned}D^{\texttt{3-1D-GAF},}_{IS} = d_{IS}\sqrt{\frac{16\pi}{5}}\Bigg[A_a\left\langle Y_{2,0}\right\rangle_{\texttt{3-1D-GAF}} \\ +\sqrt{\frac{3}{8}}A_r\left(\left\langle Y_{2,-2}\right\rangle_{\texttt{3-1D-GAF}}+\left\langle Y_{2,2}\right\rangle_{\texttt{3-1D-GAF}}\right)\Bigg]\end{aligned} \tag{6.10}$$

$$\begin{aligned}D^{\texttt{3-0D-GAF}}_{IS} = d_{IS}\sqrt{\frac{16\pi}{5}}\Bigg[A_a\left\langle Y_{2,0}\right\rangle_{\texttt{3-0D-GAF}} \\ +\sqrt{\frac{3}{8}}A_r\left(\left\langle Y_{2,-2}\right\rangle_{\texttt{3-0D-GAF}}+\left\langle Y_{2,2}\right\rangle_{\texttt{3-0D-GAF}}\right)\Bigg]\end{aligned} \tag{6.11}$$

6.3 MATERIALS AND METHODS

6.3.1 *Experimental Data*

All experimental data used here came from the literature [112]. This ensemble of RDCs is the one that was use from previous GB3 `GAF` studies [167]. RDCs used in this study are summarized in Table 6.

Table 6 – Data used for GB3 analysis. Detailed experimental conditions can be found in the reference [112].

Number	Media	$^1D_{NH}$	$^1D_{C'N}$	$^1D_{C'C^\alpha}$
0	PEG/hexanol	49	52	55
1	Bicelles	49	52	51
2	Negatively charged gel	48	51	54
3	Positively charged gel	48	53	54
4	Phages 100 mM NaCl	50	53	54

6.3.2 *SF-GAF Analysis*

A complete `SF-GAF` analysis was applied using a protocol similar to the one presented in Section 4.2. The protocol was slightly modified for this study:

- Tensors previously obtained using `1D-GAF` analysis [167] were used as a starting point. Therefore steps 1-3 were skipped. The peptide planes were actively included in the tensor refinement if the dataset contained at least 20 couplings. For the tensors scaling using the "K_A" protocol, not only $^1D_{NH}$, but also other couplings ($^1D_{C'C^\alpha}$ and $^1D_{C'N}$) were used for indirect analysis.

- Optimization of $\sigma_{\alpha,av}$ and $\sigma_{\beta,av}$ was achieved using a grid search. Starting from zero values, $\sigma_{\alpha,av}$ and $\sigma_{\beta,av}$ were incrementally increased until reaching a global χ^2 minimum. The first search was made with 1° steps. Results were refined using 0.1° steps. This protocol allows easier $\sigma_{\alpha,av}$ and $\sigma_{\beta,av}$ estimation and simultaneous optimization of orientations and dynamic amplitudes, leading to a more precise determination of these values.

6.3.3 *Simulated Data*

Some simulated data were produced to test the `LS-GAF` models. Data were simulated using the β-sheet structure of GB3. Tensors used are the five tensors optimized in the `SF-GAF` procedure (see above). Noise is added using a Gaussian noise generator and standard deviations were fixed by the weight optimized in the `SF-GAF` analysis (typically between 0.3 and 1 Hz for $^1D_{NH}$). The following datasets were simulated:

1. A completely static description. Noise was added using Guassian widths of one (dataset S1) and two standard deviations (S2).

2. `3D-GAF` motion. The orientations of the reorientation axes were fixed using the GB3 β-sheet structure. Amplitudes of reorientation were set using a random distribution and restricted to the following domains:

   $$\sigma_\alpha \in [0;3], \qquad \sigma_\beta \in [0;8] \qquad \text{and} \qquad \sigma_\gamma \in [0;20]$$

   Data were simulated in the absence of noise (LR0) and with a Gaussian widths of one standard deviation (LR1).

3. Same procedure as 2, but `3D-GAF` motion assumed to be identical for all the planes with $\sigma_\alpha = 6.26, \sigma_\beta = 7.36$ and $\sigma_\gamma = 11.34$. No noise was added in (LF0) and noise was added in (LF1).

4. A `3-0D-GAF` is assumed. The orientation of the motion is arbitrarily fixed. Amplitudes of common reorientations are fixed to $\sigma_\delta = 6.26, \sigma_\epsilon = 7.36$ and $\sigma_\zeta = 11.34$. Data were simulated without (GF0) and with noise (GF1).

5. A `3-3D-GAF` is assumed. The global reorientation is described using common reorientation amplitudes of $\sigma_\delta = 6.26, \sigma_\epsilon = 7.36$ and $\sigma_\zeta = 11.34$ with arbitrary fixed orientations. Local motion is simulated with directions fixed by the β-sheet structure and random amplitudes of reorientation uniformly distribute within:

   $$\sigma_\alpha \in [0;3], \qquad \sigma_\beta \in [0;5] \qquad \text{and} \qquad \sigma_\gamma \in [0;8]$$

   Data were simulated in the absence (LG0) and presence of noise (LG1)

6.3.4 *GAF Collective Motions*

Collective motion was investigated using the `3-0D-GAF` model. Here, a fragment of the GB3 protein is supposed to share common motional behavior. This approach was applied to the β-sheet (residues 9-13, 18-23, 47-51 and

56-60) and the α-helix (residues 27-42). Orientations of the reorientation axes and associated amplitudes of motion were determined using orientations optimized in the `SF-GAF` analysis. Obtained results will be noted for orientation axis $\mathcal{A}^0_\delta$, $\mathcal{A}^0_\epsilon$ and $\mathcal{A}^0_\zeta$ and for amplitudes of reorientations σ^0_δ, σ^0_ϵ and σ^0_ζ.

This approach was first of all made with the entire RDC dataset. Then two RDCs per peptide plane were randomly removed from the dataset. The analysis was redone with this new dataset and the indirect χ^2 (χ^2 of the unused data) were estimated on the basis of the obtained results. This procedure was repeated ten times and χ^2_{ind} is obtained by averaging the indirect χ^2 over these repetitions.

In order to compare results with other reasonable descriptions, a model `S-G` where the whole fragment undergoes an identical reorientation in a cone motion `S` is used.

6.3.5 *Global and Local Motion Determination*

As simultaneous determination of all the three amplitudes of shared motion (σ_δ, σ_ϵ and σ_ζ) and the appropriate number (one or three) of individual amplitude of reorientations are computationally very time consuming, a sequential determination was used.

1. Local orientational properties ($\mathcal{A}_\alpha$, $\mathcal{A}_\beta$ and $\mathcal{A}_\gamma$) are determined using the orientations obtained in the `SF-GAF` analysis.

2. Global axes of reorientation ($\mathcal{A}_\delta$, $\mathcal{A}_\epsilon$ and $\mathcal{A}_\zeta$) are fixed to that determined in the `3-0D-GAF` protocol ($\mathcal{A}^0_\delta$, $\mathcal{A}^0_\epsilon$ and $\mathcal{A}^0_\zeta$).

3. Global angles of reorientation (σ_δ, σ_ϵ and σ_ζ) are determined using the `3-0D-GAF` results. The three values are simultaneously scaled by A_S in order to sweep between $(0, 0, 0)$ and $(1.5\,\sigma^0_\delta, 1.5\,\sigma^0_\epsilon, 1.5\,\sigma^0_\zeta)$. Therefore, amplitudes of the motion are changed without modifying the relative distribution of the motion.

4. Local amplitudes of reorientation according to `1D-GAF` or `3D-GAF` model are determined.

Similarly to the previous step, this approach was first made with the entire dataset and then complemented with ten indirect analyses (the same data are removed as in the previous analysis) and indirect χ^2_{ind} is averaged over those ten repetitions. For `3-1D-GAF` model, a comparison is made with

a model where the whole β-sheet is submitted to an isotropic S motion and each peptide plane experiences its own local motion, i.e. a γ-motion (S-1D-GAF).

6.4 RESULTS AND DISCUSSION

6.4.1 *Tensor Determination*

The relative properties of tensors were accurately defined by the previous GAF studies of GB3. Therefore, the relative tensor orientation and rhombicity remain globally unchanged during this new optimization. However, the use of the K_A scaling protocol reveals the impact of tensor magnitude and the absorption of the dynamic component of the tensor eigenvalues (Figure 32). The similarity between these results and those obtained from simulated data from GB3 (Figure 17) and experimental data from Ubiquitin (Figure 16) is striking, indicating a possible general nature of the SF-GAF approach. The minimum found according to the 3D-GAF indirect analysis will be considered as a quantitative determination of the magnitude of the studied tensors. Final tensors can be found in Table 7.

Table 7 – Alignment tensors determined during GB3 analysis.

Tensor	A_a (10^{-4})	A_r (10^{-4})	α (°)	β(°)	γ(°)
0	−8.55	−1.68	93.15	79.24	173.03
1	−15.63	−3.52	−80.27	124.30	−7.99
2	11.15	4.78	109.82	99.25	122.67
3	9.85	6.71	−51.37	70.53	54.01
4	13.46	1.14	−75.38	90.81	−35.16

6.4.2 *Local Dynamics Analysis*

Averaged angles of reorientation $\sigma_{\alpha,av}$ and $\sigma_{\beta,av}$ were determined to be equal to respectively 4.6 and 10.0 °. 28 peptides planes are modeled using M-I, 19 with M-II and 7 with M-I. No evident correlation between the repartition of the models and the secondary structure were found. Average values of reorientation amplitudes eventually obtained are:

$$\langle\sigma_\alpha\rangle = 5.03\,^\circ, \quad \langle\sigma_\beta\rangle = 8.39\,^\circ \quad \text{and} \quad \langle\sigma_\gamma\rangle = 11.70\,^\circ \tag{6.12}$$

Averaged amplitudes of reorientation are on the same order as those obtained in previous 3D-GAF studies [167].

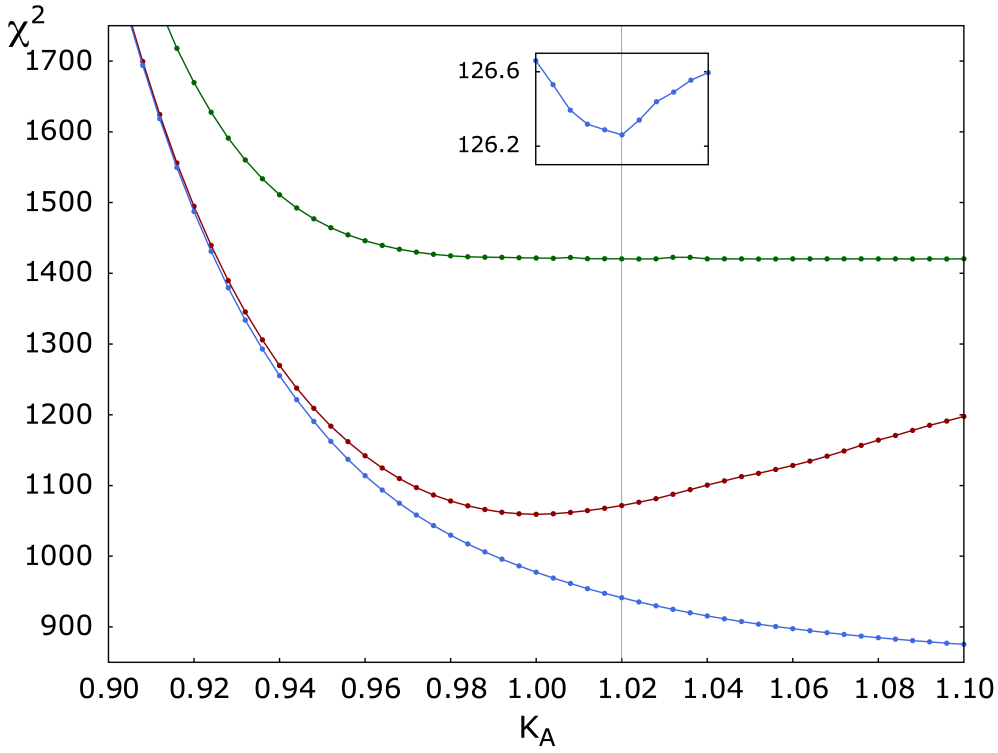

Figure 32 – GB3 experimental data. Effect of the alignment tensor scaling on direct data reproduction χ^2 according to model `S` (green), `1D-GAF` (red), `3D-GAF` (blue). The scaling is applied according to K_A. The value $K_A = 1$ corresponds to the tensors obtained after `1D-GAF` optimization. The inset corresponds to indirect data reproduction according to `3D-GAF`. Grey line corresponds to the optimal tensors.

Local dynamic amplitudes and their comparison to previous results of the GB3 `3D-GAF` analysis and accelerated molecular dynamics of GB3 [181] are shown in Figure 33. The corresponding numerical values can be found in Tables 19 and 20 in Annexe C. Comparing the `SF-GAF` and `3D-GAF` approaches leads to essentially identical results in terms of the S^2_{NH} order parameter distribution. The main difference resides in the level of dynamic amplitudes: the `SF-GAF` approach leads to slightly higher level of dynamics, except for a few exceptions. This shift can be explained by the slightly higher alignment tensor eigenvalues for the `SF-GAF` analysis, which thus allows for the presence of a little more dynamics. Comparison with the absolute order parameters derived from this AMD is less robust than for Ubiquitin, as the AMD was applied in a more primitive way, with an *ad hoc* adjustment of the level of acceleration in order to match the lowest $S^2_{NH,AMD}$ with the lowest $S^2_{NH,GAF}$ of a `3D-GAF` analysis. Moreover the final step using standard MD was not applied [167].

Figure 34 presents N_i-H^N_i order parameters and γ-motion obtained from the `SF-GAF` analysis and the previous `3D-GAF` analysis. The general pattern

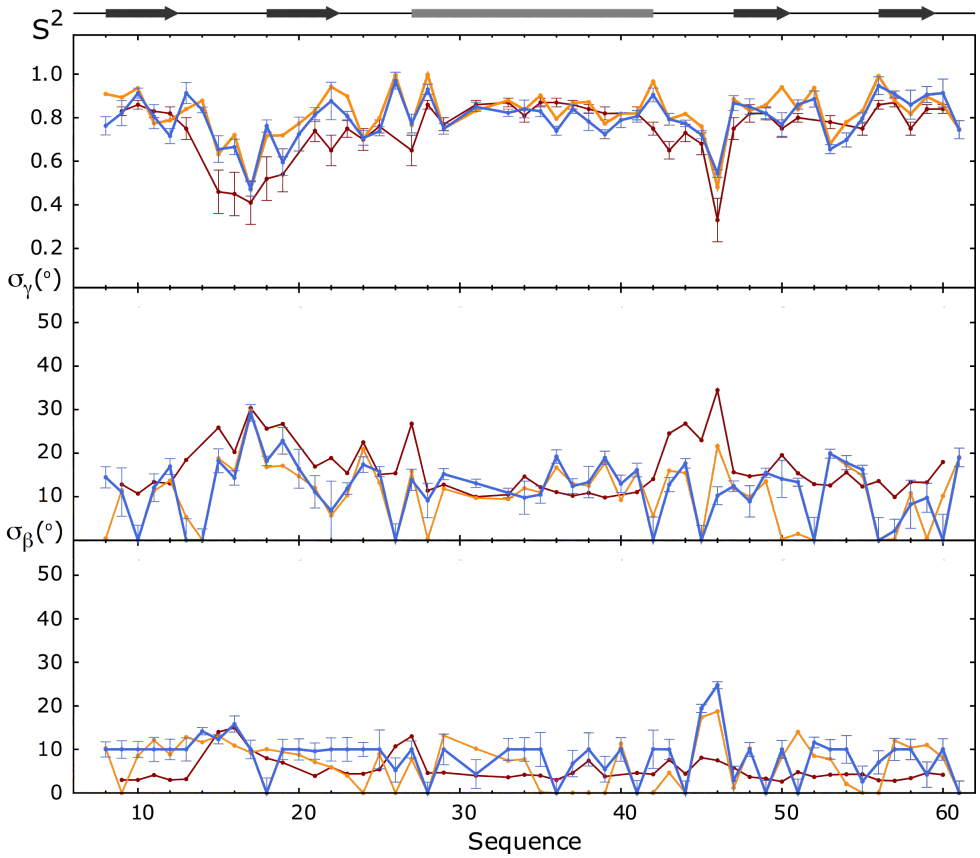

Figure 33 – Local GB3 dynamics obtained through SF-GAF analysis. N_i-H_i^N order parameters (upper panel) and amplitudes of reorientation for γ-motion (central panel) and β-motion (lower panel) derived from SF-GAF analysis (blue), previous 3D-GAF analysis (orange) and Accelerated Molecular Dynamics simulation (red). AMD slower order parameter was set to be equal to the lowest 3D-GAFN_i-H_i^N order parameter.

is similar with some local differences, as shown from Figure 33. It is worth noting that the main motion revealed by Normal Mode Analysis [189] corresponds to a "breathing" motion leading to maximal amplitudes of motion in the first N-terminal loop (the one on the upper left part of Figure 34). The `SF-GAF` analysis seems to confirm the presence of such a motion as the dynamics in this region increases as compared to the `3D-GAF` analysis.

6.4.3 *Simulated Data: Identification of Collective Motion*

A `3-0D-GAF` analysis was performed using all the simulated datasets. The three amplitudes of global reorientation (σ_δ, σ_ϵ and σ_ζ) and the three corresponding ($\mathcal{A}_\delta$, $\mathcal{A}_\epsilon$ and $\mathcal{A}_\zeta$) axes were determined simultaneously. No local dynamics were assumed. Motional amplitudes are shown in Table 8.

Table 8 – Testing `3-0D-GAF` model on GB3 simulated data. Values of the global reorientation angle fitted using `3-0D-GAF` models with various simulated datasets. Details of each simulated data can be found in the Materials and Methods section. Briefly the number in the dataset name indicates the level of noise, the F or R letter indicates whether the three local amplitudes of reorientation are fixed (F) or randomly distributed between zero and the indicate value (R). For clarity the three global amplitudes of reorientation were permuted in order to be sorted by decreasing value.

	Fitted values			Values used to simulate data						
Data	σ_δ	σ_ϵ	σ_ζ	σ_δ	σ_ϵ	σ_ζ		σ_α	σ_β	σ_γ
S1	2.6	1.7	0.0	0.0	0.0	0.0	F	0.0	0.0	0.0
S2	3.5	2.3	0.0	0.0	0.0	0.0	F	0.0	0.0	0.0
LR0	2.0	1.7	1.6	0.0	0.0	0.0	R	3.0	8.0	20.0
LR1	3.5	2.3	0.0	0.0	0.0	0.0	R	3.0	8.0	20.0
LF0	10.4	9.2	7.7	0.0	0.0	0.0	F	11.3	7.4	6.3
LF1	10.5	9.5	6.8	0.0	0.0	0.0	F	11.3	7.4	6.3
GF0	11.3	7.4	6.3	11.3	7.4	6.3	F	0.0	0.0	0.0
GF1	11.9	7.9	7.2	11.3	7.4	6.3	F	0.0	0.0	0.0
LG0	13.3	9.4	6.9	11.3	7.4	6.3	R	2.0	5.0	8.0
LG1	12.5	10.0	5.3	11.3	7.4	6.3	R	2.0	5.0	8.0

Concerning the static descriptions (S1 and S2), no significant dynamics are fitted using the `3-0D-GAF` model. Small reorientation angles are fitted in order to best accommodate the random noise. This gives an initial estimation of the `3-0D-GAF` model noise sensitivity.

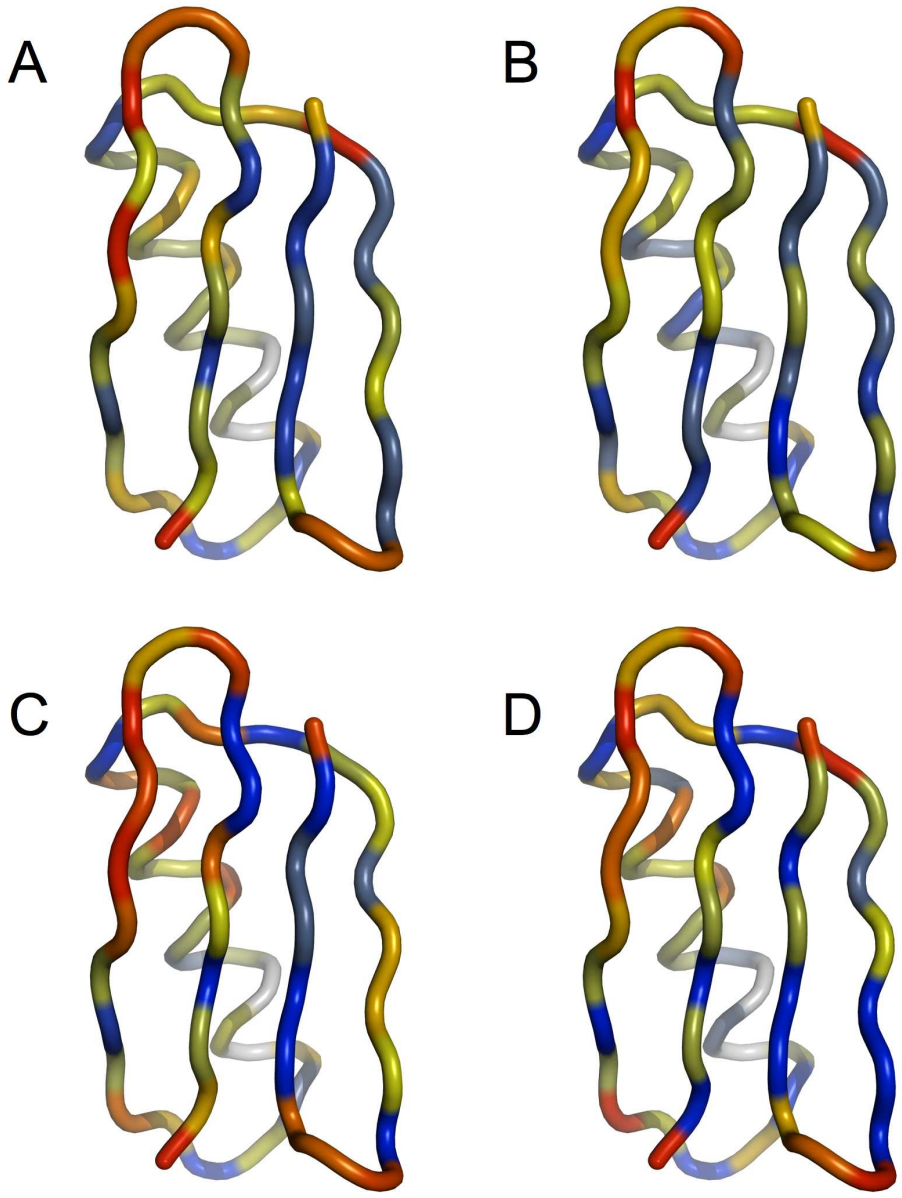

Figure 34 – Local GB3 dynamics obtained through `SF-GAF` and previous `3D-GAF` analysis. N_i-H_i^N order parameters (A and B) and amplitudes of reorientation for γ-motions (C and D). `SF-GAF` results are on the left (A and C) and previous `3D-GAF` ones on the right (B and D). Color scale for γ-motions from dark red ($\sigma_\gamma > 20\,°$) to dark blue ($\sigma_\gamma < 6\,°$) via green, yellow, orange. Color scale N_i-H_i^N order parameters from dark blue ($S^2_{NH,GAF} = 1.0$) to dark red ($S^2_{NH,GAF} = 0.5$) via green, yellow, orange. Grey: not determined. The presented structure was determined using DYNAMIC-MECCANO [167].

If randomly distributed 3D-GAF (LR0 and LR1) are used, no collective motion is detected within a few degrees of accuracy. If a 3D-GAF model with constant motional amplitude is used, significant amplitudes can be fitted. Compared to the three introduced angles, the range of the three fitted values is narrower. Considering a local amplitude distribution with similar amplitudes of reorientation, the effect of the 3D-GAF can be crudely represented as essentially isotropic. The effect is therefore to scale down corresponding simulated RDCs[1], even if this global reorientation is not able to correctly reproduce the detail of the simulated RDCs, it does reproduce the correct range for the calculated RDCs. If a broad distribution of local motion amplitudes is present no overall scaling effect can be found and no collective motion is extracted. If a constant local motional model is present, analyzing them with only collective motions will lead to a factor that simply scales down RDCs with a loss of information about local anisotropy.

In the case where only global motion is introduced in the simulation of the data (GF0 and GF1), the global reorientation is perfectly determined in the absence of noise and within an accuracy smaller than $1\,^{\circ}$ in the presence of reasonable levels of noise.

If a global motion is introduced in the presence of randomly distributed local motion, the global motion can be estimated within an accuracy of $2\,^{\circ}$.

These simulations indicate that the 3-0D-GAF model is able to extract correct amplitudes of collective motion in the presence of experimental noise and in the presence of randomly distribute local dynamics. 3-0D-GAF does not overinterpret the random fluctuation due to local dynamics or Gaussian noise as global motion. Nevertheless, the presence of identically distributed motion along the sequence can be partially interpreted as global motion. Analysis of experimental data with 3-0D-GAF model is expected to reveal a collective motion if this corresponds to the "reality" of the motion experienced by the different planes but with a slight overestimation of the amplitudes of reorientation in the presence of local motion.

6.4.4 *Anisotropic Collective Motion Analysis*

3-0D-GAF analysis was made on two different fragments of GB3 protein, the β-sheet and the α-helix. Amplitudes of collective reorientation are shown in Table 9.

1 This rudimentary analysis neglects all anisotropic aspects of GAF models (e.g. GAF motion can increase some RDC values of specifically oriented vectors [161]) and is just used to underline how the 3-0D-GAF model can absorb part of the local motion.

Table 9 – Amplitudes of reorientation obtained from `3-0D-GAF` analysis of GB3 α-helix and β-sheet. The three amplitudes of reorientation were sorted by decreasing values.

Fragment	σ_δ	σ_ϵ	σ_ζ
β-sheet	11.24	8.40	5.70
α-helix	10.43	9.66	3.83

The data reproduction for the whole fragments with direct and indirect analysis are presented in Table 10.

Table 10 – Data reproduction using `3-0D-GAF` or `S-G` analysis of GB3 for α-helix and β-sheet fragments. Direct χ^2 correspond to the data reproduction of all the data within the considered fragment. Indirect χ^2 are calculated only for the removed data and averaged over ten different calculations.

Fragment	χ^2(`3-0D-GAF`)	χ^2(`S-G`)	χ^2_{ind}(`3-0D-GAF`)	χ^2_{ind}(`S-G`)
β-sheet	593.03	651.31	34.96	36.14
α-helix	416.71	436.71	21.75	21.69

Direct analysis with the `3-0D-GAF` model improves data reproduction compared to that obtained with the `S-G` description. Statistical tests on direct χ^2 indicate a relevance of the use of the `3-0D-GAF` model for the β-sheet. Concerning the α-helix, the improvement is statistically valid (e.g. it passes a 5% F-test) but less important than for the β-sheet. Indirect data reproduction validates the `3-0D-GAF` improvement compared to a `S-G` analysis for the β-sheet, whereas improvement is less obvious for the α-helix. Thus the following study will focus only on the β-sheet.

The axes of reorientation of the β-sheet fragments are presented in Figure 35. Clearly, the two main axes of reorientation lie in the "plane" defined by the β-sheet. The accuracy of their determination was investigated with 100 Monte-Carlo simulations. The results, presented in Figure 36, clearly indicate a rather poor determination of those directions. Concerning the main axis, it mainly remains in the β-sheet "plane", whereas the second axis can clearly move out of the plane. Those Monte-Carlo simulations allowed the estimation of the accuracy of the three amplitudes of reorientation, leading to:

$$\sigma_\delta = 11.24 \pm 2.09, \qquad \sigma_\epsilon = 8.40 \pm 2.4 \qquad \text{and} \qquad \sigma_\zeta = 5.70 \pm 2.72. \tag{6.13}$$

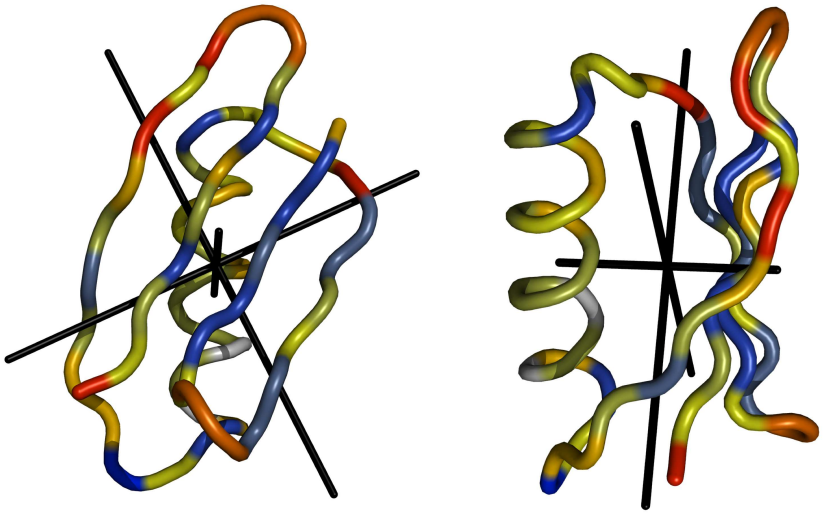

Figure 35 – Axis of reorientation obtained from `3-0D-GAF` analysis of GB3 β-sheet. The frame origin is set arbitrarily at the center of gravity of the protein. Lengths of the three axes are proportional to their corresponding amplitude of reorientation. Color scale, representing `SF-GAF` order parameter: from dark blue ($S^2_{NH,GAF} = 1.0$) to dark red ($S^2_{NH,GAF} = 0.5$) via green, yellow, orange. Grey: not determined. The presented structure was determined using DYNAMIC-MECCANO [167].

6.4.5 *Simultaneous Characterization of Anisotropic Collective and Individual Dynamics*

The simultaneous determination of local and collective dynamics using `LS-GAF` models is applied by progressively scaling the collective motion. Data reproduction according to the `3-0D-GAF`, `3-1D-GAF` and the `3-3D-GAF` models are presented in Figure 37. The arbitrary factor A_S is set to 1, at the reorientation angles obtained in the `3-0D-GAF` analysis. For A_S values above 1, the three approaches rapidly converged to identical data reproductions. For A_S values smaller than 1, the behavior of the three studies differs. For the `3-0D-GAF` analysis, the χ^2 quickly decreases from high values for small A_S to reach a minimum at the reorientation angles found in the previous step. The `3-3D-GAF` starts at a plateau value, with a small increase of the χ^2 as a function of A_S. The curvature increases until it is indistinguishable from the other two curves. Concerning `3-1D-GAF` χ^2 variation, the corresponding curve remains in between the two others, but reaches a minimum for $A_S = 0.85$, corresponding to reorientation angles of:

$$\sigma_\delta = 9.55, \qquad \sigma_\epsilon = 7.14 \qquad \text{and} \qquad \sigma_\zeta = 4.84. \tag{6.14}$$

The variation of A_S corresponds to the influence of the amount of collective motion in the GB3 β-sheet. The ordinate axis of Figure 37 corresponds to the absence collective dynamics. At this value, the three obtained χ^2 correspond

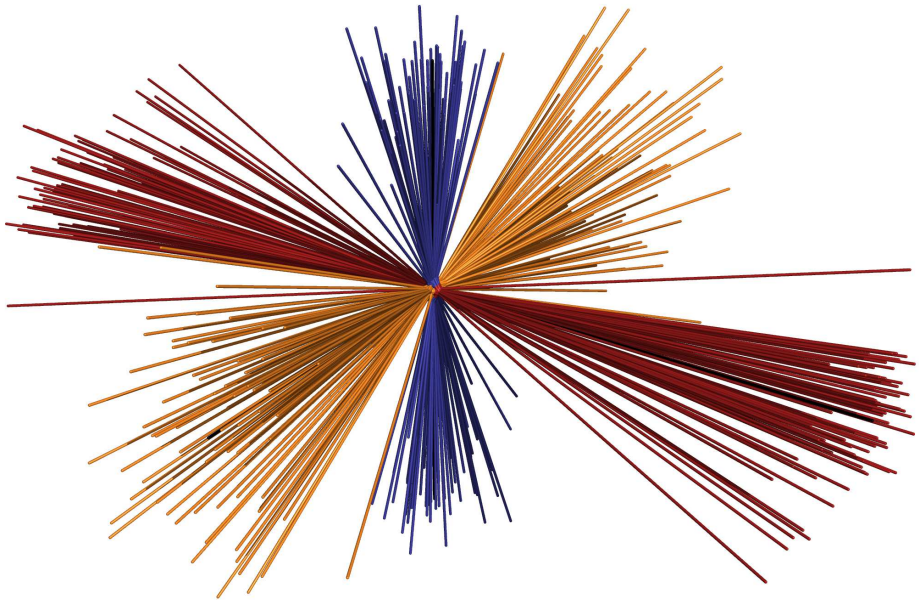

Figure 36 – Accuracy of the 3-0D-GAF analysis of GB3 β-sheet. Axis of reorientation obtained from 3-0D-GAF analysis of GB3 β-sheet through 100 Monte-Carlo simulations. Lengths of the three axes are proportional to their corresponding amplitude of reorientation. Red, orange and blue axis correspond the main, the intermediate and the smaller axis of reorientation. Black axes correspond to those experimentally determined.

in decreasing order to static, γ-1D-GAF and 3D-GAF, all measured with the same dynamically optimized tensors. For the highest values of A_S, a very large collective motion is imposed. The perfect convergence of the three curves indicates that this too-high common dynamics does not leave any room for a local motion, as both locally static and dynamic models converge to the same description.

For the 3-0D-GAF model, the evolution of the data reproduction as a function of A_S is simple: the χ^2 value decreases until it reaches an optimal value. Further increase of collective motion makes the data reproduction worse.

Concerning 3-3D-GAF motion, the interpretation is quite clear: the best data reproduction is obtained in the absence of collective motion. The presence of small amplitude collective motion does not significantly affect the data reproduction, but with an increase of amplitude, the quality of the data reproduction starts to decrease. This was confirmed by indirect analysis (data not shown). This indicates that the data does not justify more GAF reorientations than the three of the 3D-GAF model. In other words, it underlines the fact that the 3D-GAF motion is able to describe all the RDC-detectable dynamics present in the system.

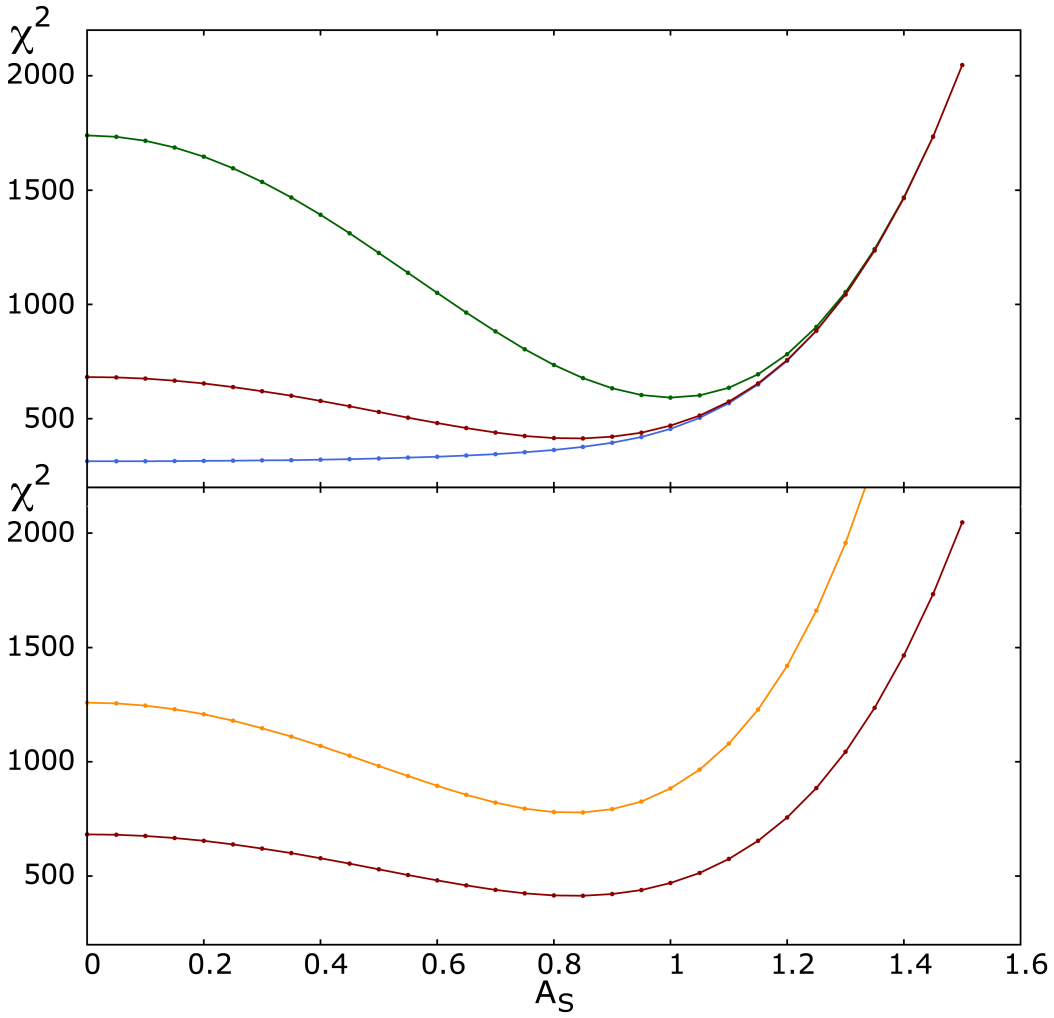

Figure 37 – Data reproduction of GB3 β-sheet RDCs according to the different `LS-GAF` models. Upper panel direct analysis: `3-0D-GAF` (green), `3-1D-GAF` (red) and `3-3D-GAF` (blue) as a function of A_S a scaling factor common to the three axis of reorientation find in the `3-0D-GAF` analysis. Lower panel direct analysis (red) and indirect analysis (orange) using `3-1D-GAF` model. For clarity the indirect χ^2 as been multiplied by a factor of 5. $A_S = 1$ corresponds to the angles determined by `3-0D-GAF` analysis.

The data reproduction according to the 1D-GAF model optimally reproduces the data at the value of $A_S = 0.85$. The relevance of this minimum was checked and confirmed using indirect analysis (see Figure 37). This model allows only γ-like local dynamics. Therefore all the dynamics present in the β-sheet cannot be interpreted with this single motion. By progressively introducing collective dynamics, the residual dynamics can be interpreted using this common motion. The optimal combination of the two motions is found for a shared motion smaller than the one obtained using only global motion (3-0D-GAF model) as both local and global dynamics have to be expressed to give a proper description of the dynamics of the β-sheet. This minimum does not correspond to a better model than a direct 3D-GAF analysis in terms of data reproduction. This is normal as the 3D-GAF has more degrees of freedom, but it differently interprets the dynamics present in the system. Here α- and β-motions that independently "break" tetrahedric junctions are removed and replaced by a common motion.

The statistical relevance of this description was tested by comparing data reproduction with the 3-1D-GAF analysis and a model where the entire β-sheet feels an isotropic S motion and each peptide plane undergoes a γ-motion (S-1D-GAF). The comparison to this model should determine whether the 3-1D-GAF model has any relevance concerning the global motion of the β-sheet, as using an overall isotropic reorientation corresponds to scaling down the alignment tensor. AIC or F-test analysis concludes that the 3-1D-GAF motion is more relevant than the S-1D-GAF description. The present results shows that a better alternative can be found by invoking a general GAF motion of the β-sheet.

Nevertheless, statistical comparison with 3D-GAF, where all peptide planes are allowed to undergo three local reorientations, indicates that the 3D-GAF remains much more likely than the 3-1D-GAF. Therefore the 3-1D-GAF cannot be considered as an alternative to better reproduce data, it just proposes a different way to interpret part of the dynamics.

6.4.6 *Order Parameters Obtained Using Simultaneous Local and Collective Descriptions*

Order parameters derived from local and collective motions are summarized in Table 21 (Annexe C.4) and Figure 38. Order parameters obtained by comparing 3-1D-GAF and 3D-GAF results are extremely similar. N_i-H_i^N order parameters S^2_{NH} are indistinguishable within experimental error. The C^α_{i-1}-C'_{i-1} S^2_{CC} order parameters are also in good agreement, with some small differences at some particular sites. S^2_{CN} and S^2_{CH} order parameter exhibit intermediate behavior due to their relative orientation in the plane.

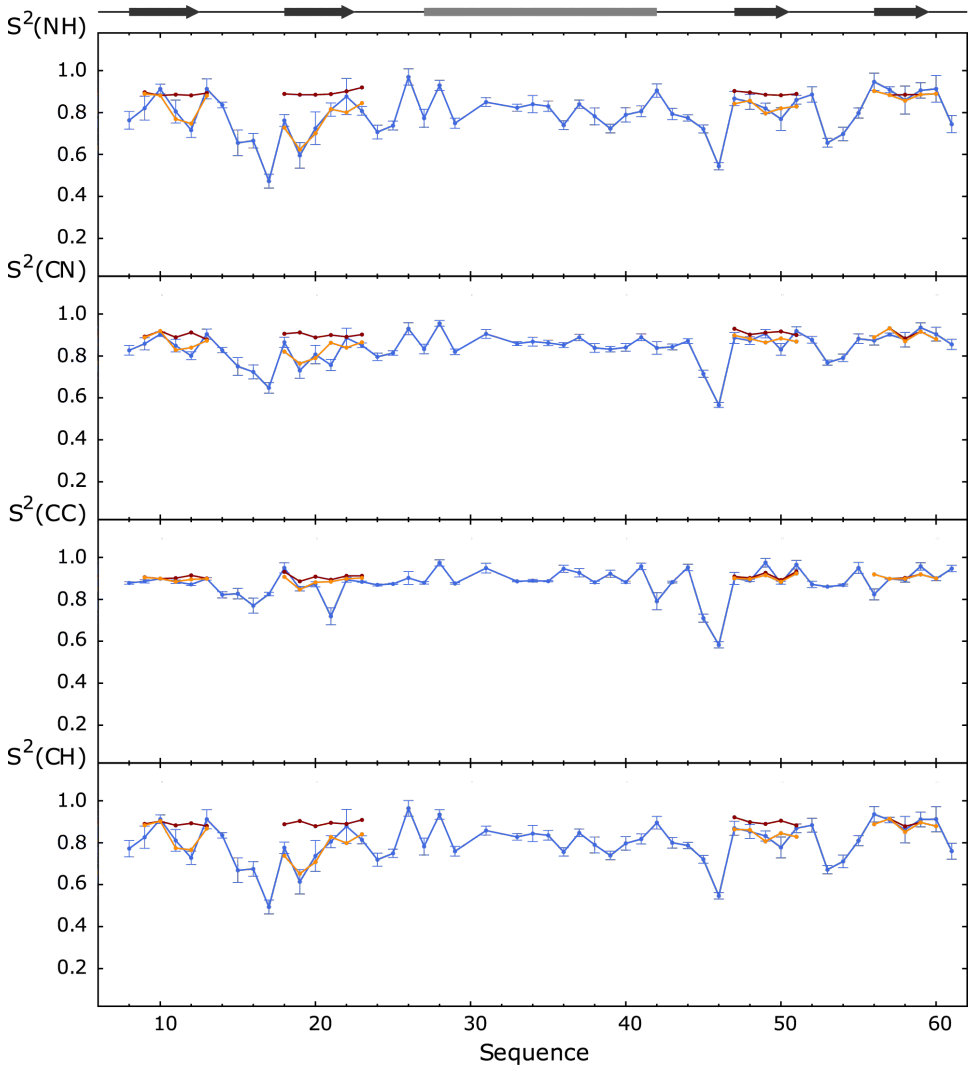

Figure 38 – Order parameters obtained through `3-1D-GAF` analysis for GB3 β-sheet: order parameters derived from the shared dynamics only (red points), from both local and shared dynamics (orange points) and compare to the order parameters derived from the `3D-GAF` analysis (blue). From top to bottom: N_i-H_i^N, C'_{i-1}-N_i, C^{α}_{i-1}-C'_{i-1} and C'_{i-1}-H_i^N order parameters

γ-motion obtained in the β-sheet using `SF-GAF` and `3-1D-GAF` model are presented on the GB3 structure in Figure 39. Clearly the amplitudes of γ-motion obtained for the `3-1D-GAF` analysis are smaller, but the relative distribution is globally similar, and a bit more homogeneous. The reduction of the range of γ-motion in the `3-1D-GAF` analysis is coherent with the fact that this model interprets part of the dynamics in terms of collective motion. The similar distribution of motion is interesting because previous GB3 studies [164] revealed the presence of correlated γ-motion within the β-sheet. The presence of a similar pattern indicates that the `3-1D-GAF` motion still allows for the existence of such a correlated motion. This motion therefore appears more as a real γ-correlated motion than a `GAF` global β-sheet correlated motion that appears through γ-motion correlation.

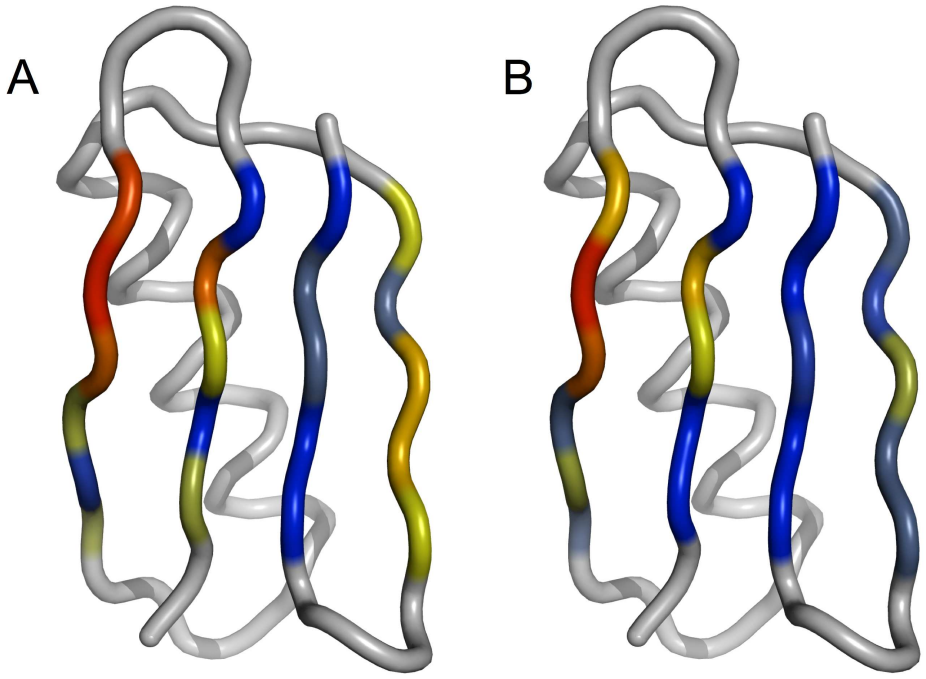

Figure 39 – γ-motions σ_γ obtained using `SF-GAF` (left) and `3-1D-GAF` (right) analysis for GB3 β-sheet. Color scale from dark red ($\sigma_\gamma > 20\,°$) to dark blue ($\sigma_\gamma < 6\,°$) via green, yellow, orange. Grey: not determined. The presented structure was determined using DYNAMIC-MECCANO [167].

6.5 CONCLUSION

In this chapter, the dynamics of GB3 was re-investigated using the `SF-GAF` description. Alignment tensor optimization led to very similar features to those described in the `SF-GAF` analysis of Ubiquitin in Chapter 4, suggesting that the approaches are generalizable.

Following this, a series of more or less complex models were developed based on the GAF description, allowing the simultaneous characterization of local and collective dynamics, and then applied to protein GB3. In the β-sheet, the results suggest the possibility to interpret part of the observed dynamics in terms of shared motion.

Here, this collective motion does not correspond to an additional motion on top of a 3D-GAF motion, but to a possible reinterpretation of the α- and β-motion dynamics, which, if applied independently, could lead to a non physical distortion of the backbone. The obtained results do not reveal major GAF collective motion in the β-sheet. This does not mean that no collectivity or correlation are present in the motion of peptide plane in the β-sheet, but that if present, they cannot be included in a simple shared GAF motion. The basic model applied here to describe motion collectivity may not be well suited to detect collective motion in the β-sheet. For example, local amplitude determination reveals increasing dynamics when the peptide plane is close to the first N-terminal loop. A better way to characterize collectivity of such a motion may be to use common reorientation axes as presented here, but with a gradient of amplitudes allowing an increase of dynamics in the interaction site. Thus the "quenching" of the shared dynamics by the less dynamic residues may be avoided or limited. The absence of data reproduction improvement using a 3-3D-GAF model also confirms the ability of local 3D-GAF motions to describe the integrality of the dynamics present in small folded proteins.

The physical nature and the corresponding picture of this motion cannot be unequivocally obtained from such a simple approach. Here, the collective motion is expressed through common axes and amplitudes of reorientation. Whether planes move with this motion in a correlated or non-correlated manner is not described by such a model. The use of complementary data (for example $^{3}J_{C'N}$ trans-hydrogen bond scalar couplings) could validate or invalidate different hypotheses.

The aim of this chapter was to determine whether non-local descriptions of the dynamics would be possible using GAF model. Here the completely local description of protein GB3 dynamics was revisited, but the range of applications of this model could be clearly very broad. Systems with more dynamic behavior could reveal the relevance of using the LS-GAF motional model. In fact multi-domain proteins or nucleic acids, with anisotropic reorientation of the different domains, or weak complexes with nonspecific interaction, could possibly require this kind of description in order to more completely reveal their dynamic behavior.

7

SH3C FAST AND SLOW DYNAMICS STUDIES

ABSTRACT

The site specific analysis of slow dynamics in proteins has mainly been applied for two model systems for which extensive data have been very accurately measured, namely Ubiquitin and protein GB3. The aim of the chapter is to investigate a new system: an SH3 domain. This system is studied both in order to increase the level of understanding of intrinsic protein dynamics, and as a first step towards eventually characterizing and comparing the slow dynamics occurring in a protein-protein complex. For this system ^{15}N relaxation and RDCs in 15 alignment media are measured. The analysis of RDCs is achieved using three different approaches: one determining an accurate averaged structure, one based on GAF models determining simultaneously structure and dynamics of the system and one based on an ensemble selection approach, where the selection procedure combines the efficiency of a genetic algorithm and model-free tensor determination using SVD. Good convergence of the different methods is observed.

7.1 INTRODUCTION

Previous chapters focused on the characterization of conformational disorder of two proteins, Ubiquitin and protein GB3, for which data were available in the literature. Nevertheless the understanding of slow timescale motions in biological system requires the investigation of as many systems as possible, to determine to which extent conclusions drawn from those studies can be generalized. In particular Ubiquitin and protein GB3 have similar folds, and it will be of interest to extend the study to differently assembled proteins.

Here the investigated system is an SH3 domain, the CD2AP SH3-C (see Annexe B). One of the interests of this system come from its interaction, in a weak complex, with Ubiquitin. Thus the study of this SH3-C presents

a double interest: characterizing the dynamics of this new system and constituting the first step towards a study of the dynamics of both components of a weak complex. This last characterization is extremely important as the functioning of a living system requires an extremely complex network of molecular interactions, between receptors and signaling molecules, between nucleic acids and proteins, between substrates and enzymes... [4, 6–8, 192, 193]. Therefore characterizing the impact of a complex formation on the protein dynamics is essential. It also carries considerable importance for the determination of the importance of so-called recognition dynamics that have been postulated to be present in both Ubiquitin and protein GB3 as the flexibility of those domains in the complex remains an essential part of the thermodynamic basis of the protein interactome.

Nevertheless studying both partners separately constitutes a prerequisite for any complex dynamics study. As Ubiquitin has already been extensively studied in Chapter 4, this chapter will deal with the characterization of the structural and dynamic behavior of the second partner the CD2AP SH3-C. Results presented here and in the following chapter are not yet definitive, in the sense that these projects are still works in progress. Nevertheless interpretation of experimentally obtained data already give some reasonable results that will be presented.

7.2 MATERIALS AND METHODS

7.2.1 *NMR Spectroscopy*

All measurements were performed at 308 K on a Varian NMR Direct-Drive Systems spectrometer 600 MHz (or 800 MHz, in this case it will be mentioned) ^{1}H Larmor frequency (14.1 T or 18.8 T) equipped with tripe resonance cryo-probes. NMR spectra were processed in NMRPipe [194] and analyzed using the program Sparky [195]. Assignment was available in the literature [196].

Relaxation measurements were performed using ^{1}H-^{15}N-HSQC, ^{15}N-T_1 and T_2 relaxation measurements, {^{1}H }-^{15}N nOe experiments and CPMG (Carr, Purcell, Meiboom, Gill) relaxation dispersion were measured. Sequences details can be found in the indicated reference [197].

For ^{13}C-^{15}N-labeled protein samples RDCs measurements ($^{1}D_{NH}$, $^{1}D_{C'C^{\alpha}}$ and $^{1}D_{C'H^{N}}$) were realized using 3D BEST-type HNCO experiments [198, 199]. For ^{15}N-labeled samples $^{1}D_{NH}$ couplings were measured from 2D SOFAST-type experiments [110].

7.2.2 *RDCs Measurements*

RDCs were measured in different alignment media (see below). All samples were prepared in 92% H_2O / 8% D_2O, buffered at pH 6.0 with 50 mM sodium phosphate (NaH_2PO_4/Na_2HPO_4) with 1 mM DTT ((2S,3S)-1,4-bis-sulfanylbutane-2,3-diol) , if slightly different conditions are used the corresponding modifications will be explicitly mentioned. SH3-C protein was provided by José Luis Ortéga Roldan. All measurements were done in Shigami tubes with 250 µL samples.

Complete sets of $^1D_{NH}$, $^1D_{C'C^\alpha}$ and $^1D_{C'H^N}$ couplings were measured in the following alignment media:

1. Bicelles. A 5% mixture (w/w) of ditetradecyl-PC (1,2-di-O-tetradecyl-sn-glycero-3-phosphocholine) and dihexyl-PC (1,2-di-O-hexyl-sn- glycero-3-phosphocholine) in a 3:1 ratio was used. The two chemicals correspond to a long (ditetradecyl-PC) and a short (dihexyl-PC) amphiphile molecules [68]. The obtained bicelles present physico-chemical properties similar as DMPC/DHPC mixtures (see Section 2.2.2), but are resistant to hydrolysis allowing much better stability of the medium.

2. Bicelles CTAB 1. Similar measurements were performed with bicelles doped with CTAB (hexadecyl-trimethyl-ammonium bromide) in a ratio ditetradecyl-PC:dihexyl-PC:CTAB 30:10:1. [68, 71, 72] and measured at 800 MHz.

3. Bicelles CTAB 2. The same protocol was used using a ratio ditetradecyl-PC:dihexyl-PC:CTAB 30:10:0.5.

4. Bicelles SDS. Similarly bicelles can be charged with SDS (sodium dodecyl-sulfate) [71, 72]. The ratio used here was ditetradecyl-PC:dihexyl-PC:SDS 30:10:1.

5. PEG/Hexanol. Measurements were made in a 5% penta-ethyleneglycol monododecyl ether ($C_{12}E_5$) hexanol mixture [79].

6. Phages. Measurements were done at pH = 6.5 , with 9 mg·mL^{-1} phages Pf1 and measured at 800 MHz [75, 76].

7. Phages Salt 1. Similar measurements were done with 15 mg·mL^{-1} phages Pf1 and 150 mM NaCl, pH = 6.5 .

8. Phages Salt 2. Similar protocol with 13.1 mg·mL^{-1} phages Pf1 and 260 mM NaCl, pH = 6.5 and measured at 800 MHz.

9. Polyacrylamide Gel. Sample were prepared using a 7% polyacrylamide gel [89, 90]. Mechanical compression was realized using direct pressure of the Shigemi plunger.

$^1D_{NH}$ couplings were also measured in:

1. Bicelles CTAB 3. Bicelles doped with CTAB in a ratio ditetradecyl-PC: dihexyl-PC:CTAB 30:10:0.1 and measured at 800 MHz.

2. Bicelles CTAB 4. Bicelles doped with CTAB in a ratio ditetradecyl-PC: dihexyl-PC:CTAB 30:10:0.2 and measured at 800 MHz.

3. Bicelles CTAB 5. Bicelles doped with CTAB in a ratio ditetradecyl-PC: dihexyl-PC:CTAB 30:10:1.5 and measured at 800 MHz.

4. Bicelles SDS 2. Bicelles doped with SDS in a ratio ditetradecyl-PC: dihexyl-PC:SDS 30:10:0.5 and measured at 800 MHz.

5. Bicelles SDS 3. Bicelles doped with SDS in a ratio ditetradecyl-PC: dihexyl-PC:SDS 30:10:1 and measured at 800 MHz.

6. Purple Membranes. Purple membrane (PM) fragments were used with 150 mM NaCl. PM were added until a reasonable range of couplings were reached (±8 Hz).

7.2.3 *^{15}N Relaxation Analysis*

The analysis of the ^{15}N relaxation data was made using the software Tensor. The principle of the analysis is the following.

In a first step the diffusion tensor is estimated using R_1 and R_2 rates. In the most rigid regions the the molecule, experiencing only librational motions, ratio of R_2/R_1 is weakly sensitive to fast dynamics, and thus the estimation of the diffusion tensor can be made within a good accuracy by using the least flexible residues. These can be identified on the basis of a known structure as residues in secondary structures are expected to be more rigid or on the basis of experimentally measured nOe as measured η_{NH} is a sensitive probe of fast motions. Residues exhibiting chemical exchange should also be removed from this analysis.

The diffusion tensors can be described using three analytical models of a diffusive rotor: the isotropic, the axial symmetric and the fully anisotropic ones (see Section 1.4). Those three descriptions correspond to different levels of complexity and using a more sophisticate model has to improve

data reproduction. In order to select the most suitable model two statistical tests are successively applied:

1. χ^2-test. This test is based on Monte-Carlo simulations. Using determined parameters of the model and adding Gaussian noise, proportional to the experimental uncertainty, it is possible to generate a set of experimental data, typically few hundreds. Using those data the determination of the parameters of the model are repeated, using the same procedure as for the experimental data analysis. This will provide a distribution of data reproduction, i.e. a χ^2 distribution. At typically the 95 centile of this distribution a limiting value often called $\chi^2_{0.05}$ is fixed. This correspond to a limit of confidence: if the χ^2 obtained for the experimental data analysis is bigger than this value the model can be rejected with a confidence of 95%. This test estimates whether, considering the experimental errors and the number of parameters used, the data reproduction is bad enough be statistically rejected. It is worth noting that the Monte-Carlo simulations used for this statistical test allow the estimation of the accuracy of the obtained parameters.

2. F-test. Here more than one model may not be rejected by the χ^2-test, the selection of the most suited model among them is obtained using F-statistics, as presented in Section 4.2.7.

When the properties of the diffusion tensor are determined, the local motion analysis is realized using a model-free formalism, by using R_1, R_2 and heteronuclear nOes. As shown in Section 1.4.3, the model-free approach can be expressed in different ways depending on the nature of the internal motion [21, 24, 26]. Starting from an experimental data set it is impossible to know *a priori* which description is the most appropriate. Therefore a systematic model selection is required.

For each residue, the different models, corresponding to different spectral density functions, are applied successively:

1. the very fast internal dynamics model where S^2 is the only remaining parameter as τ_I tends to zero (see equations 1.38 and 1.39), in practice it corresponds to $\tau_I < 20\,\text{ps}$.

2. the standard two parameters Lipari-Szabo model (equations 1.36 and 1.37)

3. the model 1 with an exchange contribution

4. the model 2 with an exchange contribution

5. the extended model-free description with a fast internal motion in the infinitely fast limit $\tau_F \to 0$ (see equations 1.43 and 1.44)

The different models are applied and tested using χ^2-test. If model 1 passes the χ^2-test it is accepted. For model 2 and 3 they have to pass in addition an F-test: in case of failure, the model 1, even if not accepted in the first step is used, otherwise the considered model is accepted. Often models 4 or 5 are able to correctly reproduce data as they use three parameters and, using a dataset measured at a single field, only three experimental measurements are available (R_1, R_2 and nOe). The fit for these models is therefore necessarily essentially perfect, otherwise it is rejected. If all the model fail the model that best reproduced the data is used and the failure is indicated.

7.2.4 *GAF Analysis*

The tensor determination was performed using the same protocol as presented in the Section 4.2.6 until the weight optimization and outlier detection using the `1D-GAF` model.

Then a simultaneous determination of structure and dynamics was performed using the DYNAMIC-MECCANO approach. This method allows a sequential determination of peptide plane orientation and dynamic behavior according to the `1D-GAF` model (see Section 3.6.3).

For each plane the determination of the mean orientation and the dynamic parameter is performed sequentially. The determination is always repeated four times in order to test the three different `1D-GAF` motions and the `S` isotropic reorientation. The best fitting model is retained. In order to select the correct orientation of the studied plane compared to the previous one a weak harmonic potential centered on a idealized tetrahedral junction (111 ° for the $C'C^{\alpha}N$ angle[1]) is introduced to overcome the two fold degeneracy remaining for the orientation of a plane according to RDCs (see Section 2.5). The potential is fixed to a low value in order to not influence the orientation of the correct solution. To minimize the possibility of getting a wrong orientation, if necessary, the optimization can be done each time for example for three peptide planes simultaneously, the first optimized as previously described and the two others with a `S` motion, their "real" dynamics being optimized in the two following steps. This is mainly useful if Prolines or residues with few or no experimental RDCs are present. The use of several

1 The used value does not correspond to ideal expected angle of 109.47 ° but was optimized using the set of structures used to determined peptide plane idealized geometry (see Section 3.6.3).

residues avoids a random orientation that breaks the fold of the protein. The structure calculation starts for the N-terminal but exactly similar procedure can be applied starting from the C-terminal and progressing in the reverse way or even staring at any plane and applying the procedure on the two opposite directions.

7.2.5 SCULPTOR *Structure Determination*

SCULPTOR (Structure Calculation Using Long-range, Paramagnetic, Tensorial and Orientational Restraints) [200] is an in-house module written in the Blackledge laboratory and now incorporated into CNS (Crystallography and NMR System) [201, 202] that allows for efficient structure calculations using tensorial restraints with floating tensors. The software allows the combination of standard force fields implemented for NMR structure calculation and experimental restraints via a pseudo potential. The obtained static structural ensemble is determined in this case from the RDC refinement, using all of the measured RDCs including 10 alignment media, of an existing NMR structure (2jte) determined using nOe restraints [203].

7.2.6 ASTEROIDS-SVD

A version of the ASTEROIDS algorithm was implemented that is adapted to the selection of structural ensembles from a larger pool. The principles of evolution and selection are identical to those presented in Section 10.2.4[2]. The difference comes from the method used to calculate RDCs from the considered ensemble. Concerning unfolded systems RDCs were calculated using PALES and then introduced in the selection algorithm, whereas here the RDC calculations are performed on the fly from the ensemble of structures using an SVD based approach (see Annexe A) similar to the one used for the AMD study applied to Ubiquitin. This procedure efficiently estimates tensors and data reproduction for different type of RDCs. If necessary successive applications allow the treatment of data from different alignment media. The approach presents the advantage of not requiring further scaling factors to match experimental RDCs range as determined tensors are optimal from the SVD analysis and obtained without requiring any hypothesis. A major difference is present compared to the original ASTEROIDS: for studies of unfolded states a different tensor is used for each conformer, whereas here an optimal average tensor is determined for the whole ensemble (this is again the same logic as was applied to the accelerated MD ensembles).

2 The original version of ASTEROIDS was developed for unfolded protein analyses, therefore the detailed description of the algorithm can be found in the corresponding chapters.

Here the population of solution P was still set at 100 and the number of iterations used to achieve good convergence was set to 8000 .

The starting pool of structures was made up of 10 000 structures extracted every 100 ps of a 1 µs molecular dynamics realized by Luca Mollica (unpublished data) using the GROMACS 4.0.4 package [204] and the AMBER ff99SB force field. The protein has been simulated in isobaric-isocoric conditions (NPT), at a temperature of 300 K and a isotropic pressure of 1 bar (Berendsen barostat [183]) in a periodically repeating box with 9500 explicit water molecules. Periodic boundary conditions have been applied and electrostatics has been treated buy means of Particle Mesh Ewald method [142, 184]. The system has been equilibrated for 5 ns and then simulated without restraints for 1 µs.

7.3 RESULTS AND DISCUSSION

7.3.1 ^{15}N *Relaxation*

Measured rates can be seen in Figure 40. No dispersion were observed from the CPMG experiment (data not shown). ^{15}N relaxation rates analysis allowed, first of all to characterize the rotational diffusion tensor of SH3-C. The complete tensor characterization is given in Table 11. The obtained tensor does not exhibit important anisotropy, which is not surprising considering the near spherical shape of the molecule (see below) and presents high similarity with the inertia tensor (data not shown). The corresponding correlation time of 3.273 ns is in good agreement with a protein of this size at 35 °C. From the obtained diffusion tensor an analysis of local mobility

Table 11 – Rotational diffusion tensor for SH3-C. Tensors eigenvalues are given in $10^{-8}s^{-1}$, angles in degree.

D_{xx}	D_{yy}	D_{zz}	α	β	γ
0.459±0.09	0.501±0.10	0.589±0.14	69.86±14.78	83.51±4.25	-83.33±5.65

was performed using a model-free approach (see Section 7.2.3). Obtained order parameters can be found in Figure 45 (see bellow) were they are compared to RDCs derived dynamics. An appropriate model of dynamics was found for every site except residue 12. Most of the residues were correctly described using model 1 or 2, i.e. standard model-free models, in the fast motion limit for model 1. The presence of exchange was invoked for the N-terminal part (3-6), for clusters of residues (13, 16, 18) , (33, 35) , 50 and (57, 59). Those region mainly correspond to *a priori* flexible parts, i.e. loops or peptide chain extremity.

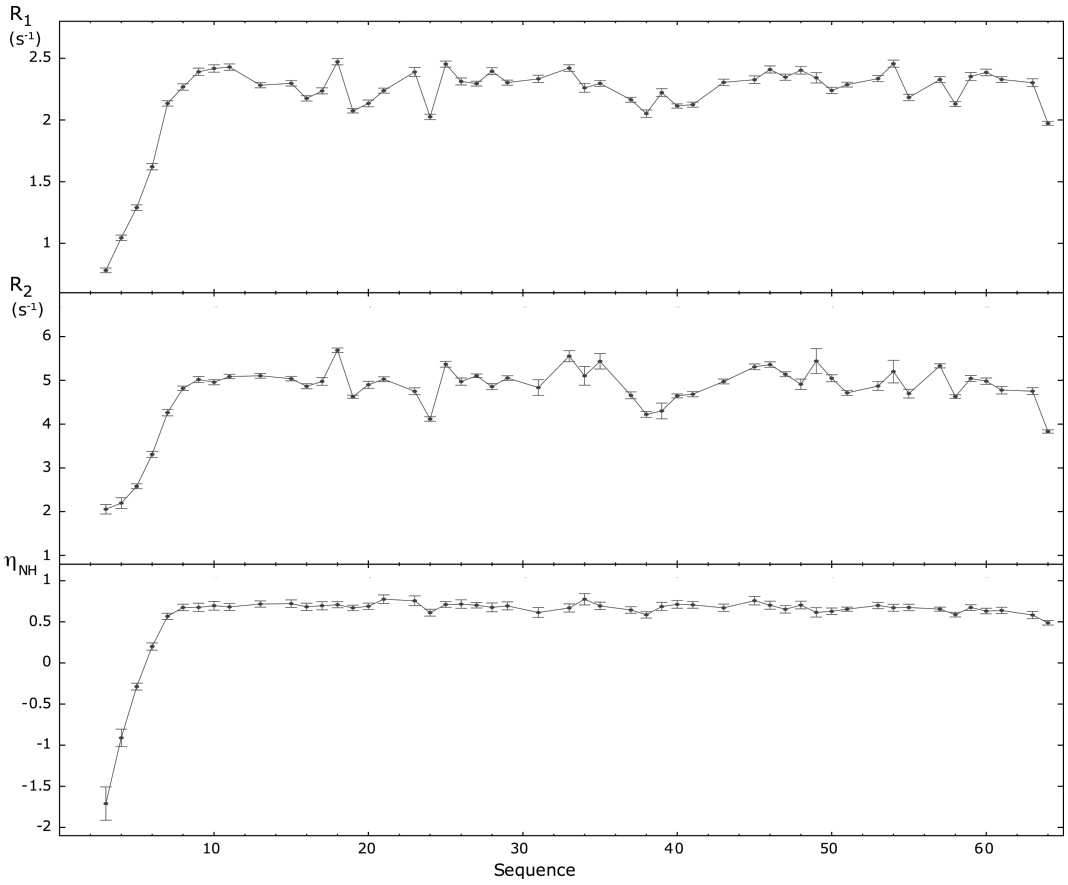

Figure 40 – ^{15}N R_1 (top), R_2 (middle) relaxation rates and $\{^1H\}$-^{15}N nOe (bottom) for SH3-C at 600 MHz ^1H Larmor frequency.

7.3.2 SECONDA *Analysis and High Resolution Structure Determination*

A SECONDA analysis (see Section 3.2.2) was performed on the set of 15 $^1D_{NH}$ datasets. By repeating the analysis for all possible combinations of 14, 13, 12, 11, 10, 9 or 8 dataset sub-ensembles, a good compromise between the consistency and the number of remaining datasets was found by selecting the 10 datasets summarized in Table 12. For this selected sub-ensemble the ratio between the fifth and the sixth eigenvalues, that characterize the self-consistency of the dataset (see Section 3.2.2), was found to be 7.9.

Structure calculations were performed using SCULPTOR. Very good convergence was achieved in the structure calculations leading to almost identical structures with an RMSD for C^α of 0.27 Å for the 10 best structures selected on the total energy: potential energies due to the force-field and the pseudo-potential of RDCs experimental restraints. The data reproduction obtained through this static description is illustrated in Figure 41. The comparison with similar structures and the fact that the planarity of the peptide plane remains well defined indicates *a priori* that the structures are reasonable

Table 12 – Alignment media selected using SECONDA analysis for SH3-C and their corresponding RDCs numbers. Details of alignment media preparation can be found in the Materials and Methods section.

Number	Media	$^1D_{NH}$	$^1D_{C'H^N}$	$^1D_{C'C^\alpha}$
0	Bicelles	56	57	56
1	Bicelles CTAB 5	60		
2	Bicelles CTAB 2	53	56	55
3	Bicelles SDS	60	60	60
4	Polyacrylamide gel	60	60	60
5	PEG/hexanol	55	55	55
6	Phages	60	60	60
7	Phages salt	58	56	60
8	Phages salt 2	55	55	55
9	Purple membrane	58		

and that the very large number of RDC restraints (around 1500) do not induce undue orientational strain or local structural deformation (94% of residues are found in most favoured regions of Ramachandran space and 4% in additionally allowed regions).

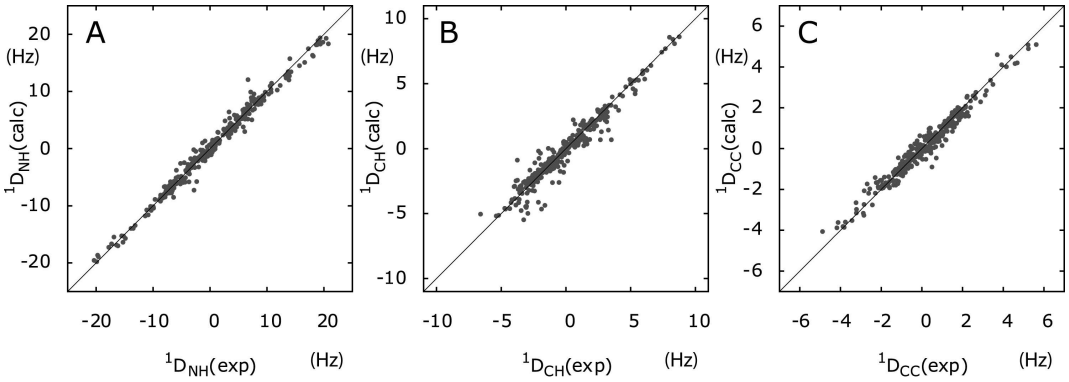

Figure 41 – Data Reproduction using a Static description. Reproduction of all (A) $^1D_{NH}$, (B) $^1D_{C'H^N}$ and (C) $^1D_{C'C^\alpha}$ couplings according to the best SCULPTOR structure.

The ten best structures of SH3-C are presented in Figure 42A. Figure 42B shows five different high resolution X-ray structures of different SH3 domains, obtained from crystal structures, in interaction with a peptide or a protein: 1ucka (1.5 Å) [205], 1u06 (1.5 Å) [206], 2ak5 (1.85 Å) [207], 2bz8 (2.0 Å) [207] and 2j60 (2.22 Å) [208].

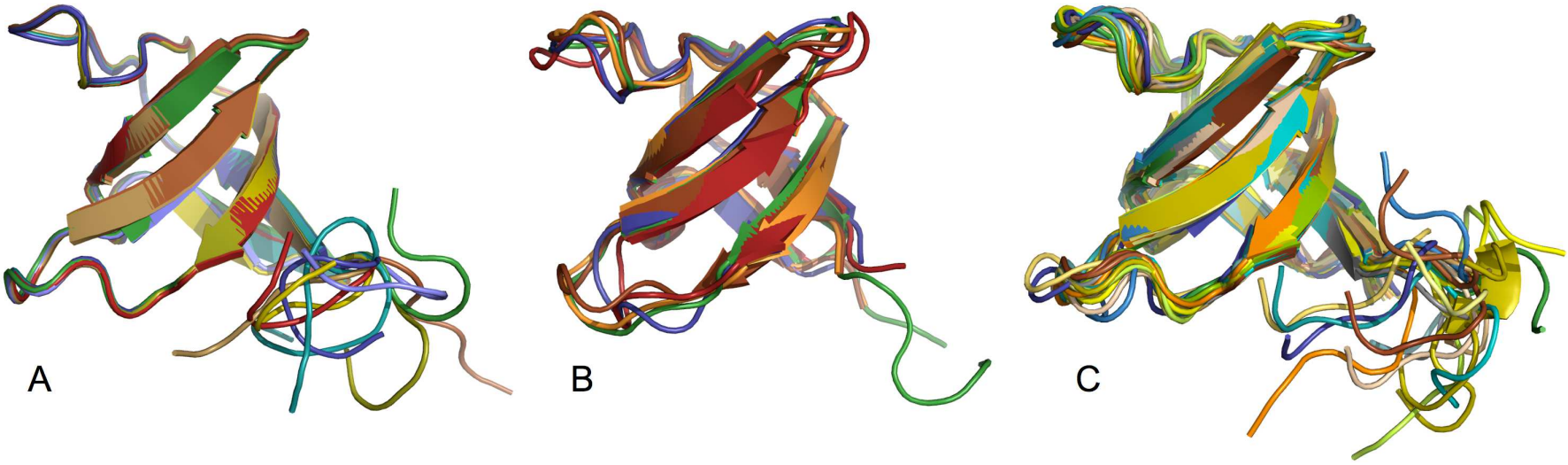

Figure 42 – Divers Representations of SH3-C. (A) 10 bests structures obtained using SCULPTOR. (B) Five different hight resolution X-ray crystal structures of various SH3 domains. (C) 20 structures from the ASTEROIDS-SVD selected ensemble.

The different SH3-C domains presented in Figure 42B does not present perfect sequence homology thus some details of the comparison may be biased by this issue. Nevertheless the comparison of the overall fold is highly similar (typical RMSD ~0.6 Å), almost identical in the β-sheet region. The main discrepancy can be found in the so-called nSrc region (loop in the bottom left part of Figure 42). In this region, which is the interaction region for the crystal structures, a high variability appears across SH3 domains. This may be due to variation in the primary sequence of the considered SH3-C or interaction with the partner with which SH3-C is co-crystalized. Considering molecular recognition mechanisms it may be interesting to determine the conformational sampling of the free SH3-C in this region to investigate the extent to which the SH3-C protein samples this conformational variability when not interacting with any partner.

7.3.3 *GAF Analysis of SH3-C Dynamics*

The protocol for tensor determination was applied on the ensemble of RDC datasets determined using the SECONDA approach. The resulting tensors, optimized using the `1D-GAF` model are presented in Table 13.

Table 13 – Alignment tensors determined during SH3-C analysis with a `1D-GAF` model.

Tensor	A_a (10^{-4})	A_r (10^{-4})	α (°)	β(°)	γ(°)
0	−5.13	−3.44	44.97	186.80	117.78
1	9.79	5.40	−187.56	99.31	−138.69
2	10.39	5.66	−185.78	98.51	−137.94
3	4.04	2.53	152.02	71.25	109.26
4	−4.17	−1.01	−178.18	107.95	−119.02
5	−13.92	−3.41	−81.60	134.79	−19.26
6	−5.31	−1.51	31.65	51.61	−116.24
7	4.68	2.61	156.75	61.09	103.13
8	4.73	2.62	157.45	60.78	102.97
9	4.04	1.01	37.01	53.47	−116.84

From these tensors, two different analyses were applied, one using DYNAMIC-MECCANO in order to obtain a simultaneous determination of structure and dynamics, according to a `1D-GAF` motion. The second analysis used a `3D-GAF` analysis as described in previous chapters. It is worth noting that results presented here are not completely definitive especially for the `3D-GAF` analysis as these experimental data have been relatively recently measured.

Nevertheless the analysis appears to be consistent and therefore provides an initial estimation of the long-timescale dynamics present in the system.

The structure obtained with the DYNAMIC-MECCANO approach is presented in Figure 43, where it is compared to the previously obtained SCULPTOR static structure. The agreement between the two structures is very good (RMSD 0.52 Å). The ability of DYNAMIC-MECCANO to determine a correct

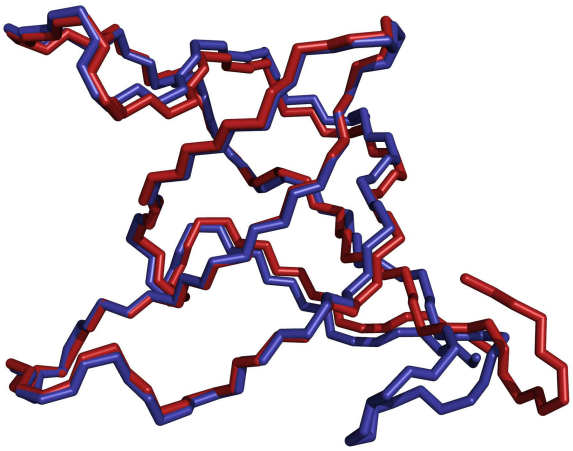

Figure 43 – Comparison of SH3-C structures obtained with SCULPTOR and DYNAMIC-MECCANO approaches: (blue) SCULPTOR structure, (red) DYNAMIC-MECCANO structure.

fold without the use of long-range information (nOe restraints, hydrogen-bonding, long-range RDCs...) underlines the accuracy with which RDCs can define orientational properties: each plane is oriented successively — using only in-plane couplings — and every error in a plane orientation can potentially propagates with dramatic effects on the protein fold. Some of the differences between the two models can be explained by the fact that one of method explicitly take into account dynamic properties, while the other (SCULPTOR refinement) ignores this component completely, and allows it to be absorbed into a floating alignment tensor. Even if the obtained structures are very similar the improvement in data reproduction clearly improve when incorporating dynamics in the description as it can be seen from Figure 44. This result indicates that the true orientational average structure determined using the DYNAMIC-MECCANO approach is very similar to the structure determined from orientationally averaged restraints, a result that is not necessarily obvious, but which is in line with results from Ubiquitin and GB3. The conformational dynamics extracted from both `1D-GAF` and `SF-GAF` description are summarized in Figure 45 and

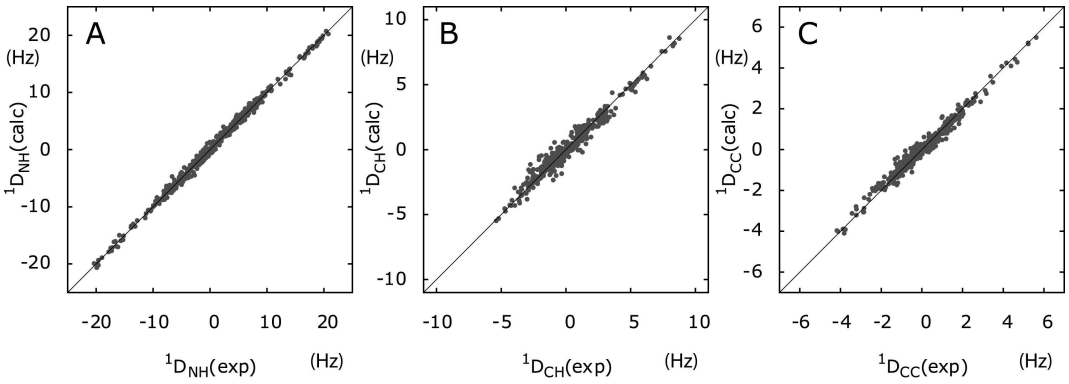

Figure 44 – Data Reproduction using a `1D-GAF` description in the DYNAMIC-MECCANO simultaneous structure and dynamic determination. Reproduction of all (A) $^1D_{NH}$, (B) $^1D_{C'H^N}$ and (C) $^1D_{C'C^\alpha}$ couplings according to a `1D-GAF` description.

represented on the DYNAMIC-MECCANO SH3-C structure in Figure 46. The order parameters obtained with the two different `GAF` analyses are found to be very similar. The `3D-GAF` exhibits slightly more dynamics due to the possibility of using three different reorientations. Comparison of `GAF` order parameters $S^2_{NH,GAF}$ and those determined from ^{15}N relaxation $S^2_{NH,REL}$ again leads to similar profiles, and to similar observations as previously described for Ubiquitin and GB3. In structured regions the two analysis revealed similar level of dynamics whereas in loop regions 15-20, 35-40 and 49-53 the `GAF` models tended to exhibit more dynamics. In the region 10-15 the $S^2_{NH,REL}$ appeared slightly lower than those of the `GAF` analysis but this may be due to non-optimal refinement of the alignment tensors (by analogy to the Ubiquitin and protein GB3 analysis the `3D-GAF` optimized tensors are expected to be slightly higher). Moreover Monte-Carlo based estimation of the uncertainties of the `GAF` obtained parameters has still to be performed to determine whether the differences are significant. Interestingly the 35-40 region, which exhibit slow dynamics, corresponds to the nSrc region, where interaction with biological partners occurs.

7.3.4 *Ensemble Description of SH3-C Dynamics*

As shown in previous chapters, an ensemble based description of the dynamics can be highly complementary to the `GAF` analysis. In this case an ensemble selection approach was used using ASTEROIDS-SVD to obtain a sub-ensemble representative of the experimentally measured RDCs from a very long MD simulation.

These approach has already been partially tested on simulated data (data not shown) and further testing procedures are in progress. Using simulated

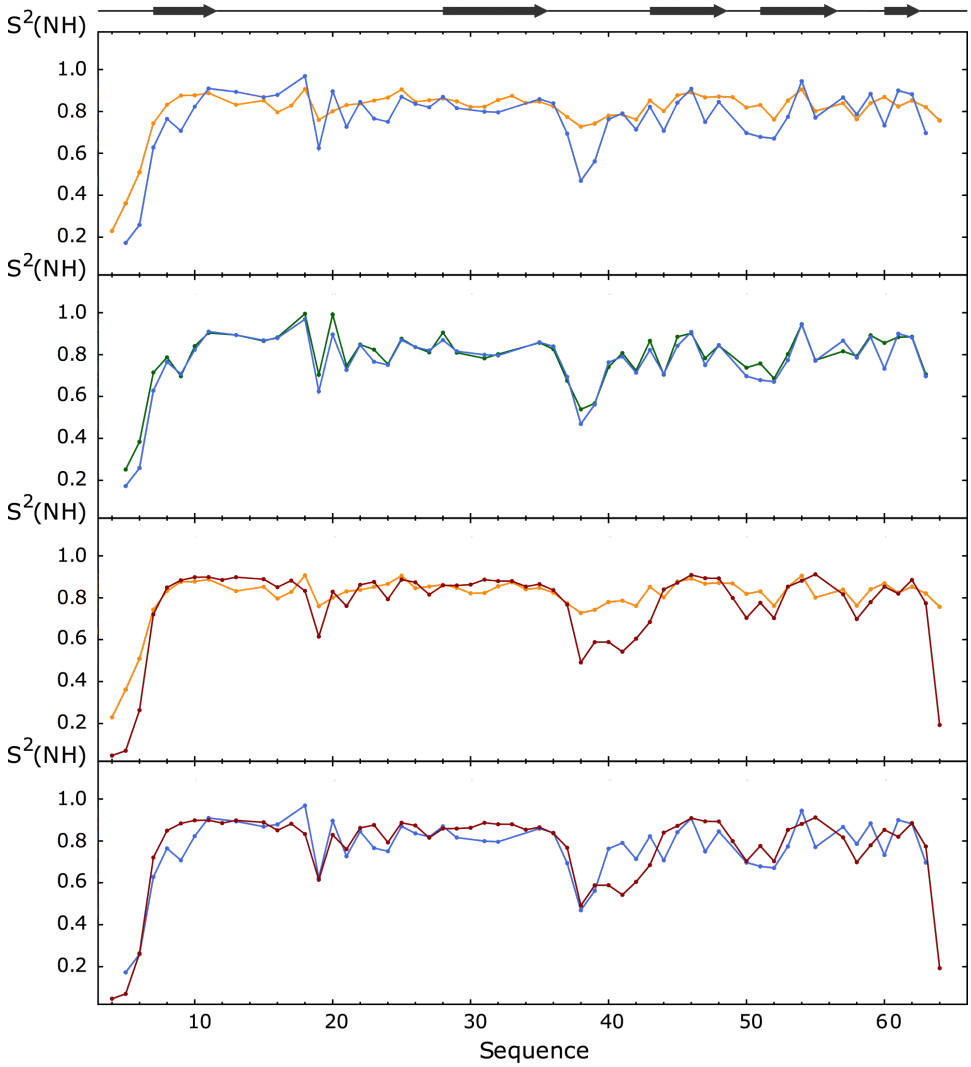

Figure 45 – N_i-H_i^N order parameters S^2_{NH} obtained for SH3-C from ^{15}N relaxation (orange), `1D-GAF` (green), `3D-GAF` analysis (blue) and using Asteroids-svd ensemble description (red). From top to bottom: ^{15}N relaxation and `3D-GAF`, `1D-GAF` and `3D-GAF`, ^{15}N relaxation and Asteroids-svd ensemble, `3D-GAF` and Asteroids-svd ensemble. S^2_{NH} corresponding to null γ-motion in the `3D-GAF` analysis are not presented. Dark grey arrows indicate β-sheet.

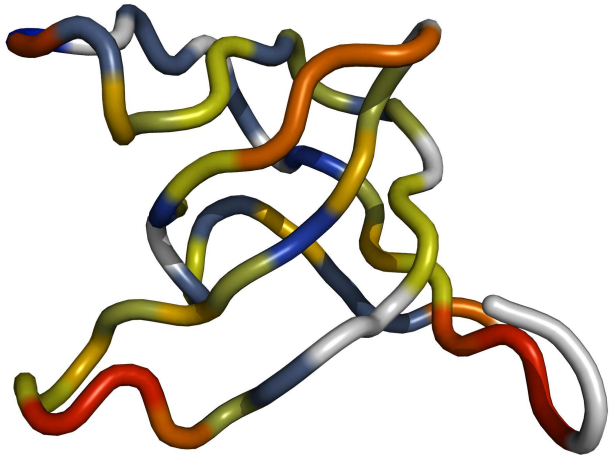

Figure 46 – N_i-H_i^N order parameters obtained for SH3-C analysis using `3D-GAF` description on its DYNAMIC-MECCANO structure. Color scale N_i-H_i^N order parameters from dark blue ($S^2_{NH,GAF} = 1.0$) to dark red ($S^2_{NH,GAF} = 0.5$) via green, yellow, orange. Grey: not determined.

data, the protocol is able to correctly recognize the conformational sampling of target ensembles on the basis of measured RDCs. Further investigation is necessary to determine the accuracy and the robustness of the approach and in particular to determine the minimal requirement of experimental data. In the current case the large volume of data measured in ten different aligment media is almost certainly sufficient. Further investigation of the influence of the initial MD sampling on the obtained results will also clarify the precision of the approach.

The analysis was made using increasing sizes of the selected sub-ensemble in order to estimate the appropriate number of structures that need to be selected. As no improvement of the data reproduction were found for ensemble size higher than 40, this size was retained for the selected ensemble. The selection procedure was repeated ten times, leading to a good convergence in terms of both data reproduction and obtained order parameters. Data reproduction obtained through this approach are presented in Figure 47.

The best ensemble of structures — the one that best reproduces the experimental data — is presented in Figure 42C and compared to the DYNAMIC-MECCANO structure in Figure 48.

Considering the comparison with the DYNAMIC-MECCANO structure the agreement is very good. Both descriptions explicitly incorporate dynamics and thus the average structure of the ASTEROIDS-SVD selected ensemble and the DYNAMIC-MECCANO structure would be expected to be in good

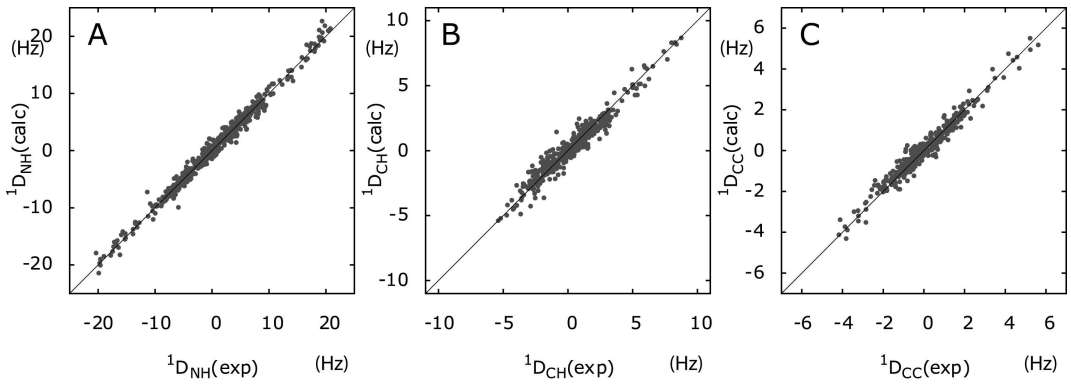

Figure 47 – Data Reproduction using ASTEROIDS-SVD description. Reproduction of all (A) $^1D_{NH}$, (B) $^1D_{C'H^N}$ and (C) $^1D_{C'C^\alpha}$ couplings according to the best ASTEROIDS-SVD selected ensemble.

agreement if the methods are appropriately applied. Here small excursions exist deriving mainly from translational differences, with the orientation of each plane being generally in good agreement in the two approaches. It is worth nothing that only 20 randomly selected structures of the ensemble are shown in the figure comparing between the ASTEROIDS-SVD ensemble and the DYNAMIC-MECCANO structure. This clearly decreases the envelope defined by the ensemble. The use of the 40-strong conformer ensemble led to a better inclusion of the DYNAMIC-MECCANO structure but also to a situation where the DYNAMIC-MECCANO structure is masked for several amino-acid, complicating the visual comparison.

The comparison of S^2_{NH} order parameters obtained from the `3D-GAF` approach and the ASTEROIDS-SVD protocol is presented in Figure 45. The overall agreement between the two RDCs based approach is qualitatively good: similar levels of dynamics are present in the two descriptions and S^2_{NH} profiles, if not identical, reveal similar patterns. The two main regions where discrepancies appear are 29-32 and 40-43. In the former the `GAF` analysis reveals slightly more dynamics. The origin of those divergences are still under investigation, but the increased dynamic for the ASTEROIDS-SVD ensemble may be due to the inability of the MD simulation to characterize the motion in this loop.

7.4 CONCLUSION

The aim of this chapter was to characterize the structural and dynamical properties of an SH3 domain. Two kinds of experiments were used to investigate this system, namely ^{15}N relaxation rates and RDC measurements.

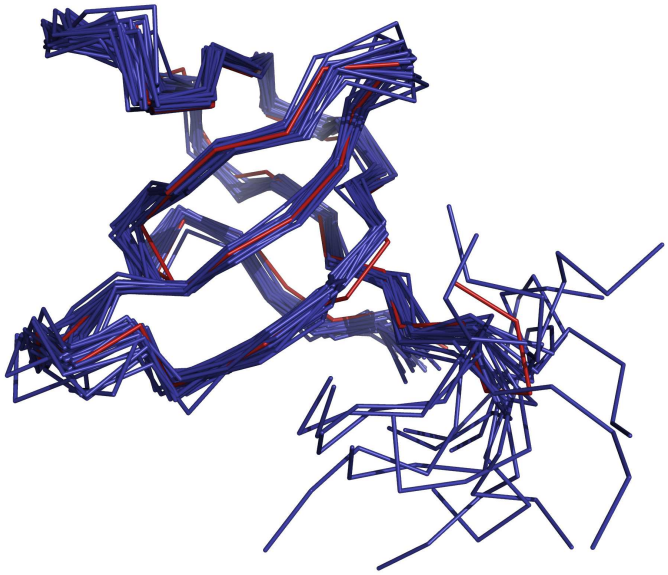

Figure 48 – Comparison for SH3-C of the ensemble of structures (blue) of the ASTEROIDS-SVD analysis and the DYNAMIC-MECCANO structure (red). For the shake of clarity only 20 conformers of the ensemble are shown.

The relaxation measurements have led to the determination of the rotational diffusion tensor and the characterization of fast dynamics of this SH3-C. The analysis mainly identifies the N-terminal part of the peptide chain as highly flexible and to detect slightly increased dynamics in the 36-44 loop region.

Concerning RDCs an extensive dataset was measured, made up 15 different alignment media and three different kind of coupling for ten of them. The self-consistency of the dataset was ensured by a SECONDA selection of a sub-ensemble of RDC datasets. The RDC analysis provided an important source of both structural and dynamic information. Even if the analysis is still in progress, some initial conclusions can already be drawn. The data were analyzed in three complementary ways: a static analysis through the SCULPTOR structure refinement using only RDCs as experimental restraints, a dynamic analysis using an ensemble selection approach based on an SVD analysis of the orientational vectors and their ability to reproduce the experimental data, and a simultaneous determination of structure and dynamics using a GAF formalism.

Concerning the structural aspect the obtained results converged to an apparently high resolution averaged structure. The SCULPTOR structure was compared to various crystal structures of SH3 domains in complex with peptides or protein. Clearly the overall fold of all those structures are nearly identical with a clear conformational variability existing for the crystal structures in the nSrc loop involved in complex formation.

Concerning the dynamic behavior of this SH3-C domain two different approaches were used and — even if not definitive — they already relatively well converged in terms of order parameters. The GAF protocol, which were developed for others studies, gave one more example of the utility of this description. The ensemble selection using ASTEROIDS-SVD appeared as an attractive alternative for the characterization of dynamics. While similar approaches already exist in the literature, this combination of a genetic algorithm and the tensor determination using SVD based method seems to provide efficient selection process. Based on completely different principles these two approaches may provide complementary information on the dynamic behavior of the studied experimental system, and allow the determination of a minimal data set that can be used for the study of future systems.

The analysis of RDCs from the GB3 and GB1 homologs and those measured in Ubiquitin have, over the last decade, stimulated extraordinary efforts to determine the true nature of dynamics occurring on timescales up to the microsecond in solution. We hope that the addition of a differently folded system, with a much lower percentage of canonical secondary structure than the previously studied systems, will provide valuable additional experimental information with which to characterize dynamics on these functionally important timescales.

The two approaches underline the presence of significant slower dynamics in the nSrc loop region, which corresponds to the interface with biological partners, and which is relatively rigid on sub-nanosecond timescales. Direct comparison with crystal structure is possibly hazardous as primary sequence of the considered SH3-C slightly differ in the region. Nevertheless concerning the question of molecular recognition it may be useful to determine whether the "conformational sampling" implied by the different crystal structures in interaction with different partners is already present in solution. In order to investigate this question the "real" conformational sampling in the complex has to be determined to allow a direct comparison between the conformational sampling of the free and the bound form of this SH3-C in interaction with Ubiquitin.

8

SH3-C UBIQUITIN WEAK COMPLEX STUDIED BY NMR RELAXATION

ABSTRACT

The study of weak protein-protein complexes is a biologically important problem, that are notoriously difficult to characterize experimentally, defying co-crystallization and presenting experimental parameters that are often confused with contributions from the free forms of the proteins. Here approaches are developed to measure ^{15}N relaxation rates for a SH3-C-Ubiquitin complex. These studies allow the estimation of the kinetics of complex formation and the derivation of accurate information about the reorientation diffusion tensor of the complex. Finally the fast dynamics in the complex can be determined and compared with dynamics in the free forms of the proteins.

8.1 INTRODUCTION

NMR is one of the most powerful tools for the study of biomolecular complexes, due to its sensitivity to protein-protein interactions with equilibrium dissociation constants varying over many orders of magnitude, from the very tightly bound, to ultra-weak complexes that can barely be detected using other biophysical techniques [209, 210]. In addition to the mapping of chemical shift changes induced by the proximity of the partner protein, and cross-relaxation derived intermolecular distance restraints, RDCs and ^{15}N spin relaxation rates [211–213] provide highly complementary orientational information defining the relative position of the partners in the complex. Although very weak protein-protein interactions (dissociation constant K_d above 10^{-4} M) are known to be important for a vast range of cellular events, such as transcription and replication, signal transduction, transient formation of encounter complexes and assembly of protein complexes, they remain the least well characterized, due to the increased difficulties encountered in isolating conformationally dependent parameters

reporting on the bound form of the proteins. Due to the weakness of the involved interactions the direct observation of the complex entity may be at best difficult and most of the time thermodynamically impossible to observe under biologically relevant conditions (see Annexe D for more detailed discussion about weak complex saturation).

The aims of this chapter is to describe a method based on standard spin relaxation measurements (^{15}N R_1, R_2 and $\{^1H\}$-^{15}N nOe) in order to quantitatively derive directly unmeasurable relaxation rates in the complex and to derive from them accurate determination of reorientation diffusion tensors. The approach consists of the measurement of relaxation rates at different molar fractions of the protein in the complex and to use this information to extrapolate the measured rates in the complex bound form. The study is applicable to pseudo two site exchange where each partner is in equilibrium between the free and the bound forms. The approach allows the estimate the kinetic constants of the equilibrium.

8.2 MATERIALS AND METHODS

8.2.1 *Protein Expression, Purification and Samples Preparation*

Protein expression and purification of both ^{15}N SH3-C and ^{15}N Ubiquitin were performed by José Luis Ortéga Roldan and Antoine Licinio. Details of expression and purification can be found in the literature [203]. The obtained Ubiquitin presents a His-tag (6 successive Histidines) on its C-terminal. Concentrations were measured by absorption measurements at 280 nm. Absorption coefficients were estimated using ProtParam algorithm at 13 980 $L \cdot cm^{-1} \cdot mol^{-1}$ for SH3-C and 1450 $L \cdot cm^{-1} \cdot mol^{-1}$ for Ubiquitin.

All samples were prepared in 92% H_2O / 8% D_2O, 50 mM sodium phosphate (NaH_2PO_4/Na_2HPO_4), 1 mM DTT at pH 6.0.

8.2.2 *NMR Spectroscopy*

All measurements were performed using similar procedures as explained in Section 7.2.1.

Two Chemical Shifts titration were performed: one for SH3-C, one for Ubiquitin. In both cases the titration started with a 0.1 mM sample of the first protein, the second being added in successive steps. Chemical Shifts perturbations were monitored using ^{1}H-^{15}N-HSQC (Heteronuclear Single Quantum Correlation) experiments.

Relaxation measurements were performed for the two proteins alone and for three mixtures of the two partners. For each point, ^{1}H-^{15}N-HSQC, ^{15}N-T_1 and T_2 relaxation measurements, $\{^{1}H\}$-^{15}N nOe experiments and CPMG relaxation dispersion were measured at 308 K at 600 MHz ^{1}H Larmor frequency. ^{1}H-^{15}N-HSQC, ^{15}N-T_1 and T_2 relaxation measurements were in addition performed at 298 K at 1 GHz ^{1}H Larmor frequency (23.5 T).

8.2.3 *Complex Kinetic and Thermodynamics*

The corresponding chemical equilibrium will be described using the formalism proposed in Annexe D. The complex formation being written as:

$$\mathbb{A} + \mathbb{B} \underset{k_{off}}{\overset{k_{on}}{\rightleftharpoons}} \mathbb{C} \qquad \text{with} \qquad K_d = \frac{[\mathbb{A}][\mathbb{B}]}{[\mathbb{C}]} = \frac{k_{off}}{k_{on}} \tag{8.1}$$

$\mathbb{A}$ being the protein of interest (either Ubiquitin or SH3-C) this equilibrium can be expressed in terms of pseudo two site exchange as:

$$\mathbb{A} \underset{k_{off}}{\overset{k_{on}[\mathbb{B}]}{\rightleftharpoons}} \mathbb{C} \qquad \text{with} \qquad k_{ex} = k_{on}[\mathbb{B}] + k_{off} \tag{8.2}$$

Considering this kinetic pathway most of the measurable quantities evolve linearly as a function of the fraction in the complex, e.g. chemical shifts, R_1 rates... but explicit contribution of exchange in the R_2 rates are also present when the site is involved in chemical shift exchange between the free and bound species. The expression of the fraction as a function of the introduced concentrations can be derived similarly to Annexe D as:

$$P_{\mathbb{A}} = \frac{K_d + [\mathbb{A}]_0 + [\mathbb{B}]_0 - \sqrt{(K_d + [\mathbb{A}]_0 + [\mathbb{B}]_0)^2 - 4[\mathbb{A}]_0[\mathbb{B}]_0}}{2[\mathbb{A}]_0} \tag{8.3}$$

8.2.4 *Chemical Shifts Titrations Analysis*

The two sets of CSs obtained with the two inverse titrations are analyzed simultaneously. As the two proteins are ^{15}N labeled and the spectrum is sufficiently well resolved, the evolution of the CSs from the two proteins can be followed in both titrations. All CS variations between the free and the bound forms are supposed to be linear as a function of the fraction of the protein present in the complex. Thus the CSs δ can be expressed as:

$$\delta_{\mathbb{A},n,i}(P_{\mathbb{A}}) = \delta^{free}_{\mathbb{A},n,i} + (\delta^{bound}_{\mathbb{A},n,i} - \delta^{free}_{\mathbb{A},n,i})P_{\mathbb{A}} \tag{8.4}$$

n corresponding to the nucleus studied (^{1}H or ^{15}N) and i to the amino-acid number.

This analysis is performed in two steps. Firstly K_d is determined during the optimization of CS changes for all residues presenting large CS variation. Then the extrapolation is made for all other residues using the previously determined value of K_d. The accuracy of the K_d estimation was made using 1000 Monte-Carlo based simulations.

8.2.5 *Relaxation Data Analysis*

Extrapolations of the R_1 rates in the complex, can be made by a similar manner as for the CS, as their variation is also expected to be linear combination of the free and bound values, thus:

$$R_{1,\mathbb{A},i}(P_{\mathbb{A}}) = R_{1,\mathbb{A},i}^{free} + (R_{1,\mathbb{A},i}^{bound} - R_{1,\mathbb{A},i}^{free})P_{\mathbb{A}} \tag{8.5}$$

For R_2 extrapolations the possibility of an exchange contribution has to be explicitly taken into account. Here the considered source of exchange is the complex formation equilibrium. Note that in one case (residue 9 of Ubiquitin) there is a clear case of conformational exchange only in the complex. For each protein, e.g. $\mathbb{A}$, the equilibrium between free and bound forms can be described as a two site exchange process (see equation 8.2), by introducing an exchange constant depending on the concentration of the second protein, e.g. $\mathbb{B}$. As K_d, k_{off}, k_{on} and k_{ex} are in this case linked the determination of K_d and k_{off} are enough to completely define this kinetic equilibrium. For a two site exchange, the analytical contribution to the transverse relaxation can be found with various expressions in the literature [29, 214, 215], leading to the following evolution of the R_2:

$$R_{2,\mathbb{A},i}(P_{\mathbb{A}}) = R_{2,\mathbb{A},i}^{free} + (R_{2,\mathbb{A},i}^{bound} - R_{2,\mathbb{A},i}^{free})P_{\mathbb{A}} + R_{ex} \tag{8.6}$$

with:

$$\begin{aligned} R_{ex} &= \frac{1}{2}\left[k_{ex} - \frac{1}{\tau_{cp}}\operatorname{argch}\left(D_+\cosh(\eta_+) - D_-\cos(\eta_-)\right)\right] \\ \eta_\pm &= \tau_{cp}\sqrt{\frac{\sqrt{\psi^2+\xi^2}\pm\psi}{2}} \qquad \text{and} \qquad D_\pm = \frac{1}{2}\left(\frac{\psi + 2\Delta\omega_N^2}{\sqrt{\psi^2+\xi^2}} \pm 1\right) \\ \xi &= 2\Delta\omega_N\left(k_{on}[\mathbb{B}] - k_{off}\right) \qquad \text{and} \qquad \psi = k_{ex}^2 - \Delta\omega_N^2 \end{aligned} \tag{8.7}$$

and $\Delta\omega_N$ the difference in ^{15}N chemical shifts between the free and the bound forms.

8.3 RESULTS AND DISCUSSION

8.3.1 *Chemical Shifts Titrations*

The effects of titration on chemical shifts are shown in Figure 49. Those two titrations were used to estimate the dissociation constant of the interaction. The obtained constant is $K_d = 0.190 \pm 0.074$ mM. The corresponding CS variations obtained for the two proteins can be seen in Figures 51 and 50.

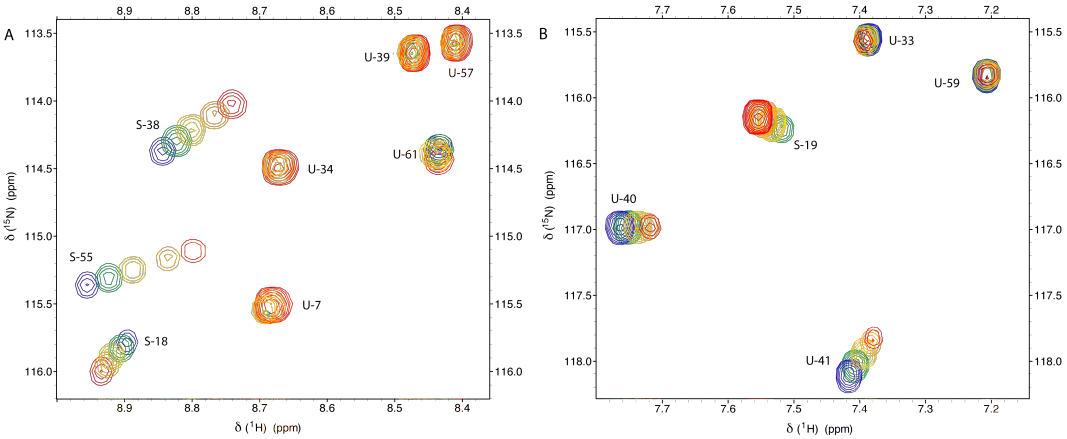

Figure 49 – Chemical Shifts Titrations of SH3-C Ubiquitin complex. Titration of SH3-C with Ubiquitin (A) and inverse titration (B). Color scale from blue to green, yellow, orange and red corresponds to an increase of the partner concentration, blue correspond to the free-form. Peaks are labeled with a letter, S for SH3-C, U for Ubiquitin and their corresponding amino-acid number.

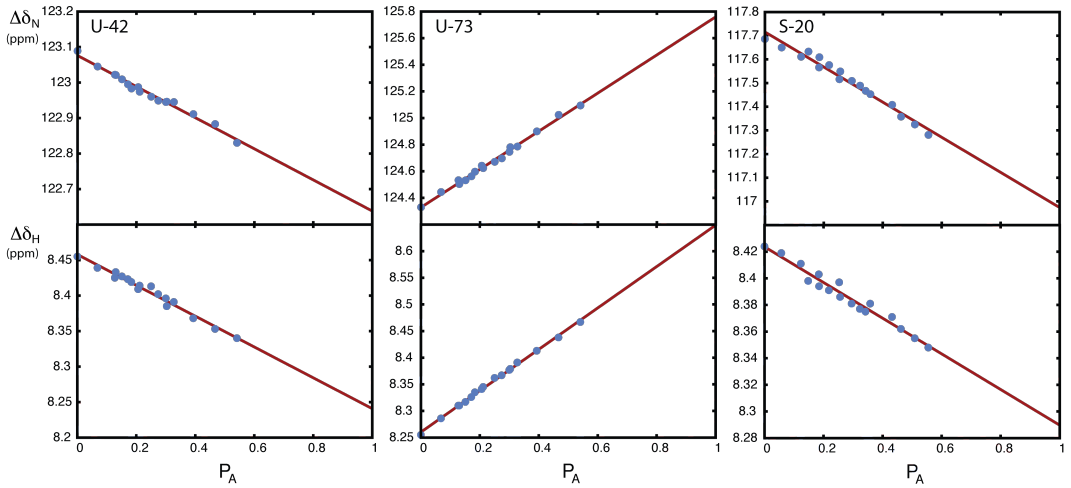

Figure 50 – Extrapolation of CSs in the SH3-C Ubiquitin complex: (blue points) experimental data and (red line) linear extrapolation. Panels are labeled with a letter, S for SH3-C, U for Ubiquitin and their corresponding amino-acid number.

The amplitudes of changes are quite small but variations are localized in

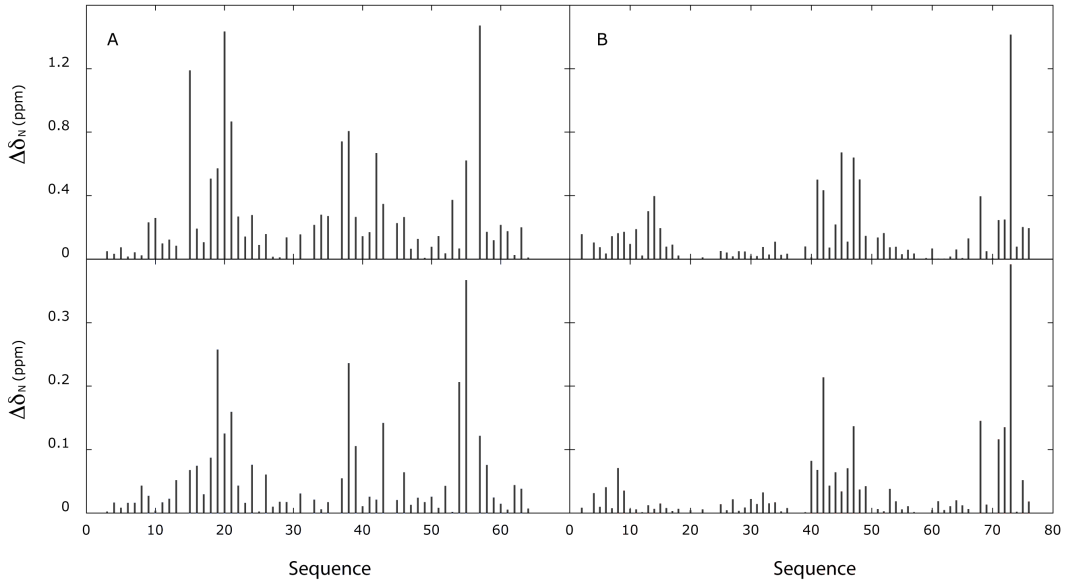

Figure 51 – Chemical Shifts variations in the SH3-C Ubiquitin complex compared to the free form: (A) SH3-C, (B) Ubiquitin. Top panel ^{15}N and bottom panel ^{1}H chemical shifts changes, both are given in absolute value.

particular regions. For SH3-C those regions are approximatively 18-23 (the RT loop), 37-39 (the nSrc loop), 42-44 (beginning of a β-strand) and 54-58. For Ubiquitin weak changes can be detected in the N-terminal part but they are mainly concentrated in the 40-50 region and the C-terminal. This distribution of CS variations is in good agreement with previous studies of the complex [196]. Moreover the formation-dissociation kinetics of the complex appears as a process in a fast exchange regime as no significant line broadening was observed during the titration.

8.3.2 *Relaxation Measurements*

The obtained ^{15}N R_1 and R_2 relaxations rates are presented in Figures 52 and 53.

By direct observation of the data some points can be immediately noticed. Clear evolution can be seen for the R_1 and R_2 rates. The overall pattern of the rates are conserved from one mixture to another, except for a few residues for which an exchange contribution clearly appears through significantly higher values of R_2 rates (e.g. residues 20 or 57 for SH3-C or residues 9 or 73 for Ubiquitin). Those contributions can be seen even more clearly from R_2 rates measured at 23.5 T (1 GHz) in Figure 54. $\{^{1}$H $\}$-^{15}N nOe measurements reveal a very similar pattern for the different mixtures (data not shown). A dynamic tail can be observed (N-terminal for SH3-C and C-terminal for

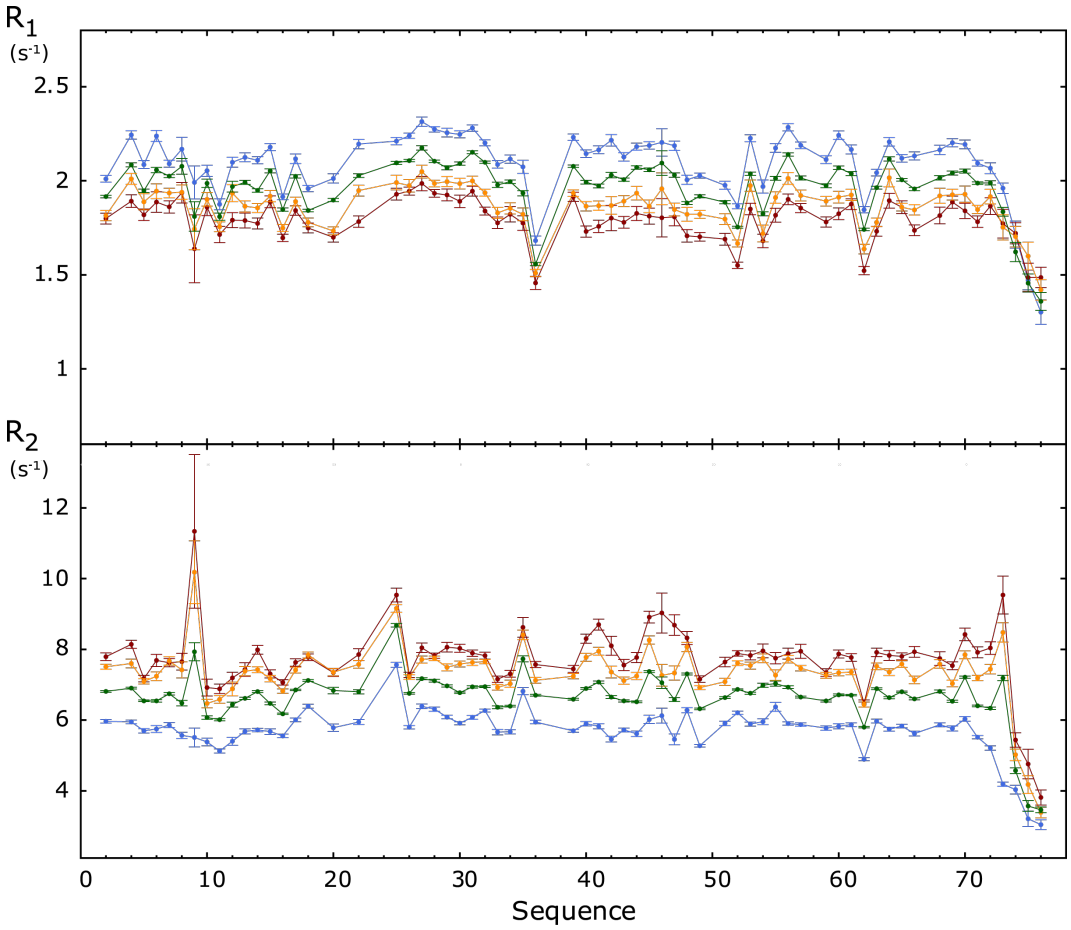

Figure 52 – ^{15}N R_1, R_2 relaxation rates (600 MHz) for Ubiquitin for samples with different protein ratios: (A) ^{15}N R_1 and (B) ^{15}N R_2. Color coding: from blue (free form) to green, yellow and red for increasing fraction in the complex.

Ubiquitin) leading to a fall in η_{NH} values, the rest of the residues exhibiting a flat profile at $\eta_{NH} = 0.6\text{-}0.7$. CMPG relaxation dispersion curves do not exhibit any significative variation for the both free forms and mixtures, thus results are not shown.

Those data may indicate that no major change in internal dynamic occurs as it would be visible from some distortion of the relaxation rates pattern. Moreover the evolution of the R_1 rates (and R_2 if no exchange is *a priori* visible) appears as qualitatively compatible with a two site exchange kinetics, i.e. evolve linearly.

8.3.3 R_1 *Rates Extrapolation*

Extrapolation of R_1 rates are straightforward as they are expected to evolve linearly as a function of the fraction in the complex. This simple evolution law allows simultaneous optimization of the fraction of the protein in the

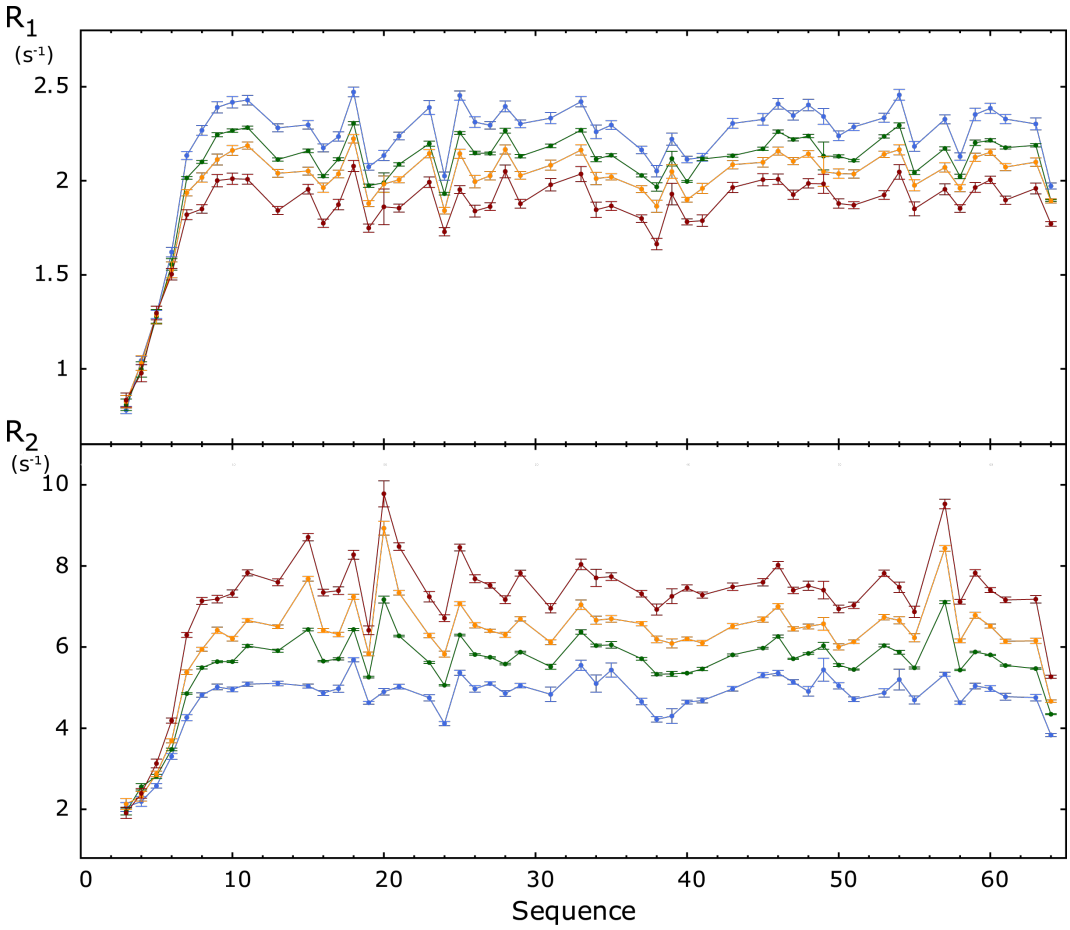

Figure 53 – ^{15}N R_1, R_2 relaxation rates (600 MHz) for SH3-C for samples with different protein ratios: (A) ^{15}N R_1, (B) ^{15}N R_2 and. Color coding: from blue (free form) to green, yellow and red for increasing fraction in the complex.

complex of both proteins. To ensure higher robustness of the fraction optimization the CSs were used too, with CSs in the free and bound form fixed to those obtained from the titration. Considering the three experimentally measured mixtures M_j (j indicating j-th experimentally realized mixture) the fractions of the two proteins evolved in an opposite way. For clarity, the fraction will be described with P_i where an increase of i corresponds to an increase of the fraction in the complex: P_i and M_j corresponds for SH3-C, but for Ubiquitin P_1 correspond to M_3 and *vice versa*. P_0 will serve to denote the free forms. The obtained fractions and typical R_1 rate extrapolations can be seen in Table 14 and Figure 55. The obtained fractions are all shifted in the same direction: smaller fraction for SH3-C and higher fraction for Ubiquitin. This revealed an initial over-estimation of the introduced Ubiquitin, and underlines the importance of the correct estimation of the concentrations of the individual proteins. Similar problems were found in previous studies of this complex [196].

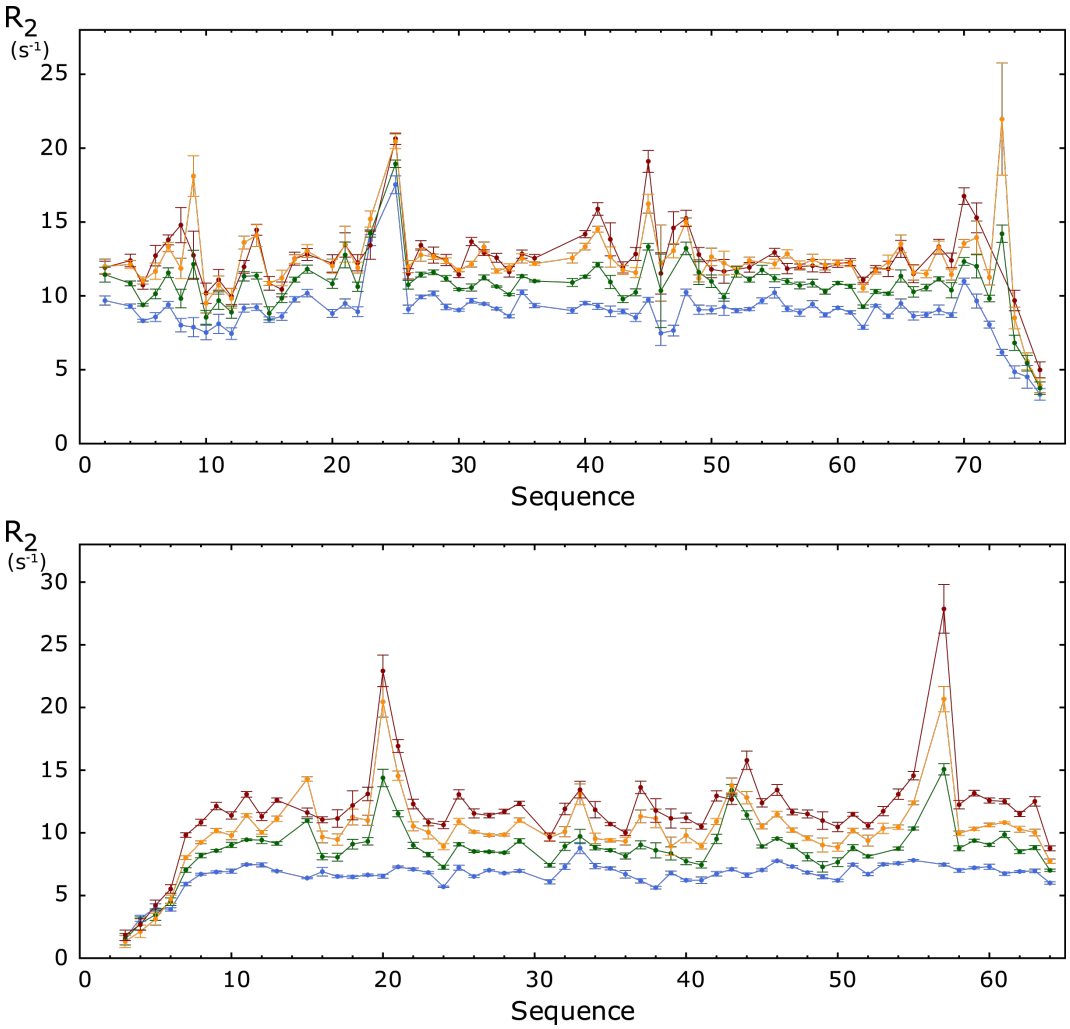

Figure 54 – ^{15}N R_1, R_2 relaxation rates (1 GHz) for Ubiquitin (A) and SH3-C (B) for samples with different protein ratios. Color coding: from blue (free form) to green, yellow and red for increasing fraction in the complex.

8.3.4 R_2 *Rates Extrapolation: First Attempts and Simulations*

Considering equations 8.6 and 8.7 and knowing expected ^{15}N chemical shifts, and differences between the free and bound forms, extrapolation of R_2 rates are in theory possible: this corresponds to the similar situation as the R_1 extrapolation with an additional common parameter k_{off} (knowing k_{off} and K_d is enough to determine k_{on} and k_{ex}, see equation 8.2), that characterize the exchange, i.e. the deviation from linearity of the R_2 rate evolutions.

Using this protocol, the extrapolation was tried, but no satisfying stability could be reached, with an apparently incompatible effects between ^{15}N chemical shifts and R_2 variations for a given kinetic constants.

Table 14 – Fraction of the protein in the complex for the mixtures of SH3-C Ubiquitin: optimized and (expected) values

	P_1		P_2		P_3	
SH3-C	0.217	(0.242)	0.360	(0.480)	0.593	(0.689)
Ubiquitin	0.284	(0.242)	0.538	(0.480)	0.695	(0.689)

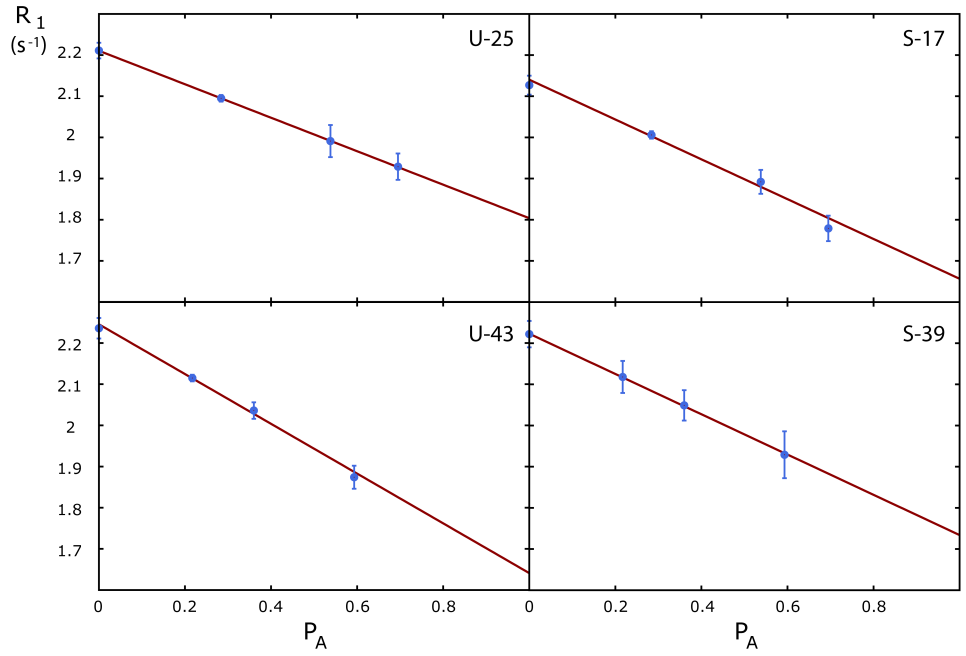

Figure 55 – Extrapolation of R_1: (blue points) experimental data and (red line) linear extrapolation. Panels are labeled with a letter, S for SH3-C, U for Ubiquitin and their corresponding amino-acid number.

In order to investigate more carefully this issue, simulations were performed on the effect of exchange on R_2 as a function of k_{off} and ^{15}N CS variations. Some of those simulations are shown in Figure 56.

The evolution of k_{off} for a given difference of ^{15}N CS between the free and the bound forms ($\Delta\omega_N$), leads to very different R_{ex} contributions. If the k_{off} is small the R_{ex} contribution appears globally linear until fraction $P_{\mathbb{A}}$ up to ~ 0.8 and an abrupt decrease occurs to reach zero for a $P_{\mathbb{A}}$ of 1. If k_{off} increases the maximal exchange contribution shifts to smaller $P_{\mathbb{A}}$ values. The maximum lies around $P_{\mathbb{A}} \sim 0.5$ for $k_{off} \sim 1000\,s^{-1}$. At this value of k_{off} the exchange contribution is maximal. Here two effects participate in this evolution: firstly the fraction of the protein in the complex and secondly the concentration of the second protein, that influences the value of k_{ex}. If k_{ex} were constant when $P_{\mathbb{A}}$ increases, the maximum of exchange would always occurs at $P_{\mathbb{A}} = 0.5$. Here the k_{ex} increases with the concentration of $\mathbb{B}$ (see

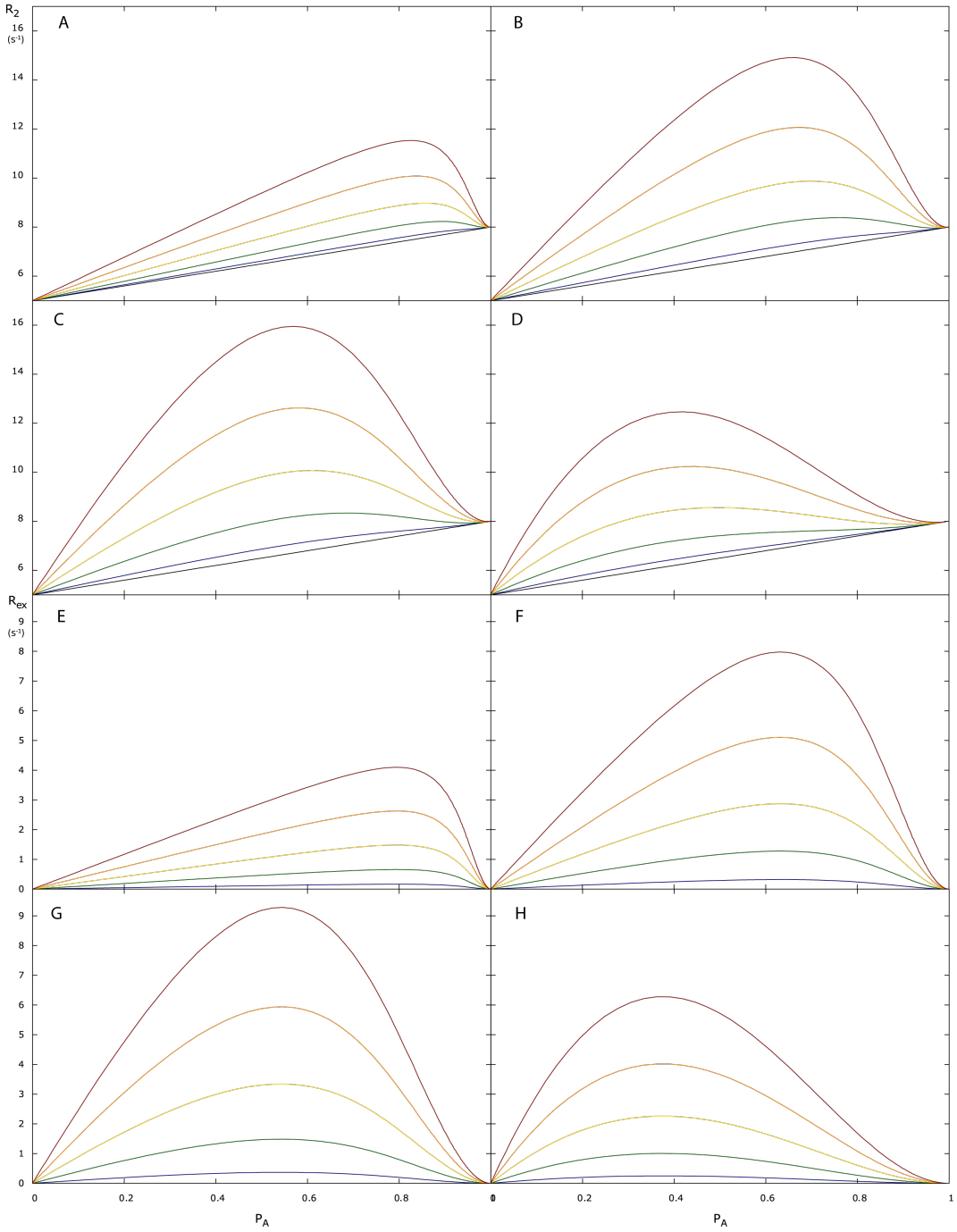

Figure 56 – Evolution of the effect of the exchange contribution R_{ex} in the R_2 measurements as a function of the fraction in the complex. Curves were simulated using equations 8.6 and 8.7. Concentrations were obtained by fixing $[\mathbb{A}] = 0.2\,mM$ and by increasing $[\mathbb{B}]$, $K_d = 0.190\,mM$. The effect of the variation of ^{15}N CS between the free and the bound forms is color coded ($\Delta\omega_N$), from 0 (black) to 1.25 ppm (red) by steps of 0.25 ppm via blue, green, yellow and orange. The R_2 rates in the free from and in the complex were set to 5 and $8\,s^{-1}$. (A)-(D) the total R_2 is represented. (E)-(H) only the exchange contribution R_{ex} is represented. (A) and (E) $k_{off} = 200\,s^{-1}$, (B) and (F) $k_{off} = 600\,s^{-1}$, (C) and (G) $k_{off} = 1000\,s^{-1}$ and (D) and (H) $k_{off} = 4000\,s^{-1}$.

equation 8.2) and thus potentially shifts the $P_{\mathbb{A}}$ value at which the maximal R_{ex} is observed.

If now we consider the R_2, the contribution due to the change in intrinsic R_2 between the free and the bound form has to be considered. If a small $\Delta\omega_N$ is considered, the R_2 will easily appear as linear at relatively high fraction for k_{off} smaller than $\sim 1000\,s^{-1}$. In this situation the possibility of discriminating for a given site the contribution due to exchange or the one due to the changes in intrinsic R_2 can be compromised. For higher values of k_{off} a clearer curvature appears allowing easier identification of the exchange contribution. We should always remember that mixtures where the bound fraction is above 0.7 are more or less impossible to attain experimentally for dissociation constants of this order or weaker, so that curvature that is defined only by a drop in the last third of the population range will not reveal useful information from this dependence.

Concerning the effect of CS, the evolution is simpler as the exchange contribution always increases if $\Delta\omega_N$ goes up, nevertheless the magnitude of this increase is modulated by the k_{off} value. Considering the range of CS variation in the titration (maximal values around 1.2 ppm) and the maximal R_2 measured in a mixture (around 10 Hz), the k_{off} as to be very big (above $\sim 5000\,s^{-1}$) or very small (below $\sim 200\,s^{-1}$) in order to fit the range of exchange contribution. Qualitatively, considering only the shape of the profile of experimentally obtained R_2 rates profiles, k_{off} can be expected to be roughly in the range of ~ 600-$1000\,s^{-1}$. However in order to fit the experimental range of R_2 rates the CS variation appears smaller to those measured in the titration.

This incompatibility seems to be the source of the previously observed extrapolation instability. In order to get around this difficulty, and obtained at least an estimation of k_{off}, the successive analyses of the ^{15}N relaxation rates in each the different mixture were realized.

8.3.5 *Evolution of Rotational Diffusion Tensors*

For each mixture and the two free forms, an analysis of the relaxation R_1 and R_2 is performed in order to estimate the effective rotational diffusion tensor of each protein in each mixture. In reality this reports on the average diffusion tensor weighted by the fraction in the complex, that has no physical meaning, but is described analytically by a real second rank tensor in the same way as for the isolated and bound diffusion tensors. This kind of analysis correlates the orientational information of the relaxation mechanisms relative to the axes of the diffusion tensor (vectors aligned with

the major axis will be relaxed as if in a larger molecule, and those aligned on the shorter axes as in a smaller molecule), and therefore requires a structural model. As no structural model is directly available for a population weighted mixture of the free and bound forms of the protein. Considering the fact that this complex is weak and that only small chemical shifts variations are observed during the titration no large structural changes are expected between the free forms and the complex. Therefore the diffusion tensor determination is performed with the structures of the free proteins (1d3z for Ubiquitin [175] and the SCULPTOR structure determined in Chapter 7 for SH3-C), but only using residues that do not exhibit significant CS changes in the titration. The two proteins are here treated completely independently.

For this tensor determination, in addition to residues exhibiting large chemical shifts changes, residues exhibiting clear exchange contribution — both are often linked — or important dynamics (e.g. C- or N- termini, flexible loop identified in the free form...) are not used. The resulting tensors are shown in Table 15. Comparing the two free forms, the diffusion tensors

Table 15 – Rotational diffusion tensor obtained for the different mixtures of SH3-C Ubiquitin. Mixture numbering increase corresponds to an higher fraction of the studied protein in the complex. P_0 correspond to the free form. Tensors eigenvalues are given in $(10^7 s^{-1})$.

	Ubiquitin			SH3-C		
Mixture	D_{xx}	D_{yy}	D_{zz}	D_{xx}	D_{yy}	D_{zz}
P_0	3.370	3.500	5.190	4.590	5.000	5.940
P_1	3.020	3.070	4.130	3.870	3.980	4.770
P_2	2.680	2.780	3.660	3.250	3.550	4.190
P_3	2.540	2.640	3.520	2.720	3.050	3.530

present significant differences. The eigenvalues of Ubiquitin diffusion tensor are smaller indicating a slower overall reorientation of the protein, which is coherent with its higher molecular weight compared to SH3-C. With the increase of the population the tensors eigenvalues tend to converge. As in the complex the two partners are expected to share the same tensor, if no differential dynamic occurs between the two proteins, this convergence correctly correlates with expected tensor evolutions observed with the increase of the protein fraction in the complex.

The obtained tensors allowed the calculation of the R_2/R_1 ratio for each studied site. The obtained results for SH3-C are presented in Figure 57. The quality of the data reproduction indicates that the structural input used

for tensor determination is well adapted, as nearly all of the site specific differences are reproduced by those calculated from the structure. Two

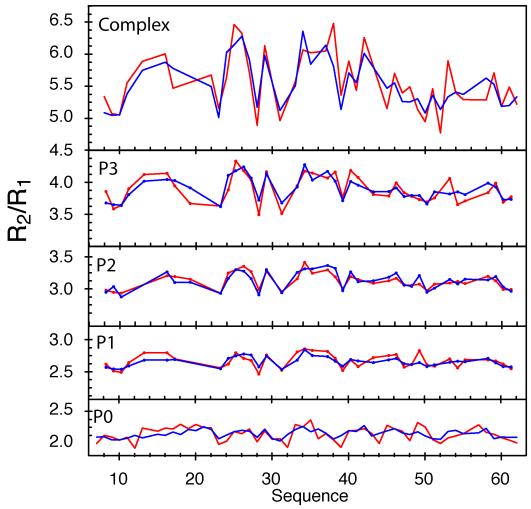

Figure 57 – R_2/R_1 ratio for SH3-C in the free form, the different mixtures and the complex. Fraction in the complex increases from bottom to top. (red) experimental data and (blue) calculated values.

phenomena can be observed from this evolution: the averaged value of the R_2/R_1 ratio increase with the complex formation and the variations of the R_2/R_1 ratio increase too. The first point is due to the longer rotational correlation time of the protein in the complex and the second to the higher anisotropy of the reorientation diffusion tensor of the complex.

8.3.6 *Kinetic Constant Estimation using Exchange Contribution Extracted from a Model-Free Analysis of Each Mixture*

The aim of this analysis was to detect the residues that exhibit significant exchange contributions due to the complex formation and to use these contributions to estimate the kinetic constants of the formation of the complex. A model-free analysis, incorporating the effective, population weighted average rotational diffusion tensor, was used to describe the contribution of overall motion to the experimentally measured relaxation rates. This allowed the local dynamics for each protein in each mixture to be estimated, but more importantly, allowed facile identification of those sites exhibiting an exchange contribution to the transverse relaxation rate. All residues presenting exchange in all mixtures and no significant exchange in the free form were identified: residues 15, 18, 20, 21, 35, 46, 57 and 59 of SH3-C and residues 54 and 73 of Ubiquitin. The result is coherent with the fact that those residues exhibited large chemical shifts changes during the titration. As residue 73 of Ubiquitin exhibits high flexibility it was not retained for this analysis.

These exchange contributions are then used to estimate the kinetic constants of the complex formation-dissociation equilibrium. As expected from the simulation, no suitable parameterization (using equation 8.7) could be found by using the CSs obtained from the titration, therefore this restraint was relaxed. Exchange rates extracted from the model-free analysis and the corresponding parameterization are shown in Figure 58. The obtained kinetic constants are:

$$k_{off} = 1564 \pm 38 s^{-1} \qquad k_{on} = (8234 \pm 199).10^3 s^{-1} mol^{-1} L \tag{8.8}$$

The obtained results are not expected to provide accurate determination of the kinetic constants, but simply to provide a reasonable range of complex formation-dissociation kinetic. The uncertainties presented here correspond to the uncertainties associated only to the last fitting step and is therefore expected to be clearly underestimated.

The discrepancy between obtained CS and those expected from the titration might be due to a more complex kinetic pathway involving processes on slow timescales, that remain invisible using relaxation measurements but that influence chemical shift values. It is worth noting that similar situation were found for the study of the interaction between another SH3-C domain and Ubiquitin using relaxation-dispersion experiments [216].

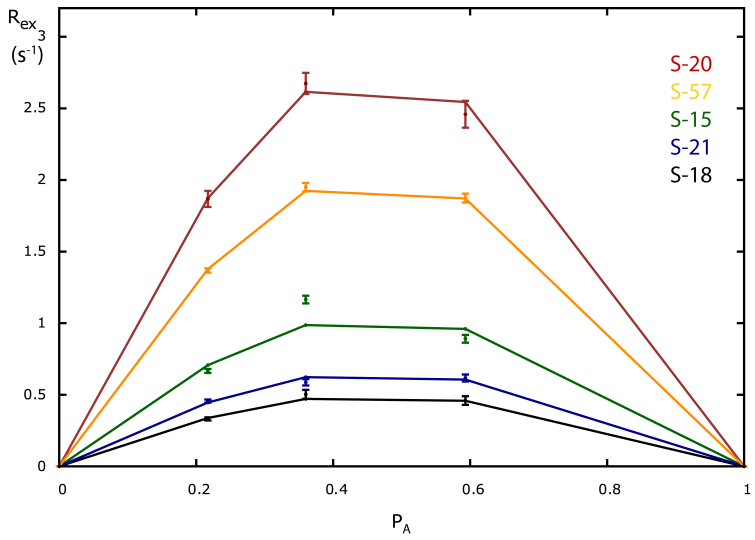

Figure 58 – Modeling of the exchange contribution in the SH3-C Ubiquitin complex formation. (Points) exchange contribution from the model-free analysis, (lines) fitted values: residue 20 (red), 57 (orange), (green) 15, (blue) 21 and (black) 18 of SH3-C.

8.3.7 R_2 *Rates Analysis using Intrinsic* R_2 *of the Complex Determined using Model-Free Analysis of the Mixtures*

During the model-free analysis the exchange contribution due to the complex formation can be determined. These contributions have been subtracted from the experimentally measured R_2, leading to a reasonably accurate estimation of the intrinsic R_2 in the mixtures. Using those values for exchanging sites and directly measured R_2 otherwise, a linear extrapolation of the R_2 rates become possible, as intrinsic R_2 are expected to evolve linearly. The obtained results are illustrated in Figure 59.

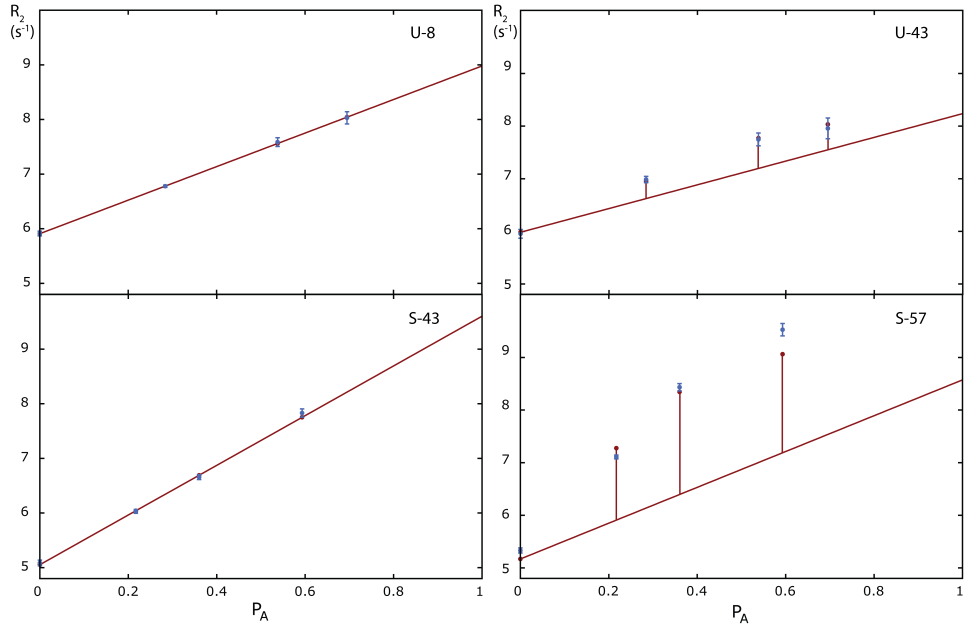

Figure 59 – Extrapolation of R_2 rates using intrinsic R_2 determined from the model-free analysis of the complex: (blue points) experimental data, (red line) linear extrapolation of intrinsic R_2 rates, (vertical red bars) exchange contributions determined during the model-free analysis and (red points) R_2 rates including the exchange contribution. Panels are labeled with a letter, S for SH3-C, U for Ubiquitin and their corresponding amino-acid number.

8.3.8 *Determination of the Diffusion Tensor of the Complex*

Due to the extrapolation procedure, residues presenting exchange contribution in the different mixtures are expected to have less accurate R_2 rate in the complex. Those residues were not used during the tensor determination, therefore avoiding any error propagation, if their determination was eventually less accurate. All residues present in flexible loops or tails are also removed.

Two different analyses were then applied: first each protein was analyzed separately as previously described for the mixtures, then the two proteins where analyzed simultaneously. In order to simplify the comparison of the analysis, the two proteins were superimposed to match the relative orientation determined from the structure of the complex determined using similarly extrapolated RDCs measured in PEG/hexanol [196]. This alignment does not influence the determination of the tensor for the individual proteins, it is simply used to facilitate the comparison of the obtained angles and the visual comparison of the tensor orientation with the shape of the complex. For the simultaneous analysis of the two proteins this orientation is on the other hand especially important as both proteins will be analyzed with the same tensor.

Results of the obtained rotational diffusion tensor are shown in Table 16 and the tensors are visually compared in Figure 60. The agreement in terms of orientations and eigenvalues is very good leading to almost identical tensors for the two proteins, using this separate determination.

Table 16 – Rotational diffusion tensor for SH3-C-Ubiquitin complex.

	$D_\perp$ $(10^{-7}s^{-1})$	$D_\parallel$ $(10^{-7}s^{-1})$	ϕ (°)	θ (°)
SH3-C	2.311±0.014	2.961±0.023	85.43±1.48	−43.77±1.26
Ubiquitin	2.289±0.013	3.054±0.024	83.25±1.36	−47.07±1.42
Complex	2.310±0.015	2.989±0.028	84.59±1.32	−44.59±1.42

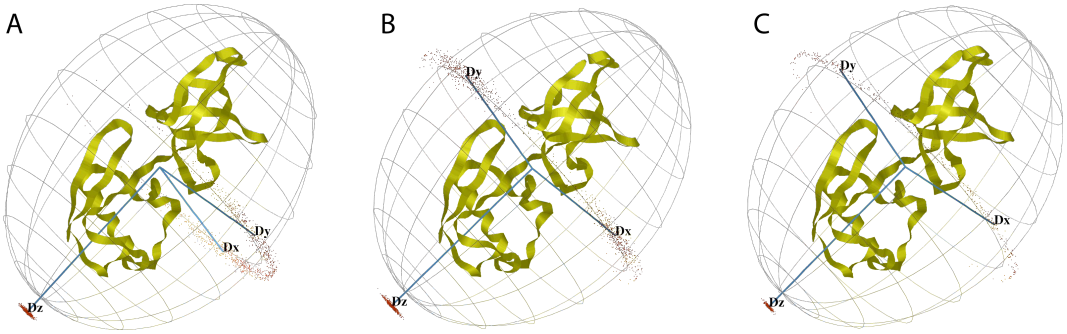

Figure 60 – Comparison of the reorientation diffusion tensor for the SH3-C-Ubiquitin complex. Tensor obtained from the analysis of SH3-C alone (A) and from the analysis of Ubiquitin alone (B) or from their simultaneous analysis (C). Points correspond to the orientations of the tensors otained using Monte-Carlo simulations.

The obtained tensors clearly present a prolate shapes. For the three analyses, the statistical analysis using χ^2-test accepted both prolate and fully asymmetric tensors. Nevertheless, using F-statistic no significant improvement

can be found using the more complex model. Thus prolate tensors are presented (see Table 16). The axial symmetry of the obtained tensors can be seen from the distribution of orientation of the reorientation diffusion tensor obtained from Monte-Carlo simulations in Figure 60 or even directly from his ellipsoidal representation. This obtained tensors seems reasonable considering the geometrical shape of the complex.

The reproductions of R_2/R_1 ratios are shown in Figure 57 and in Figure 61 as a function of α the angle between the main axis (characterized by $D_{\parallel}$) and the vector of interest. The quality of the data reproduction for both protein is

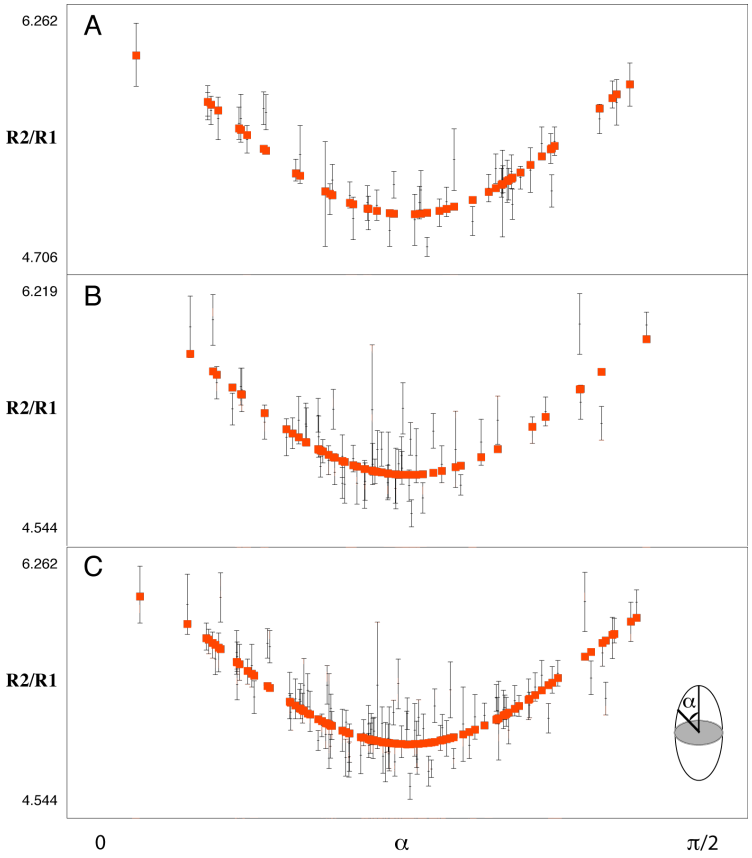

Figure 61 – R_2/R_1 data reproduction for the SH3-C-Ubiquitin complex as a function of α the angle between the main axis and the vector of interest: (A) SH3-C analyzed alone, (B) Ubiquitin analyzed alone and (C) SH3-C and Ubiquitin analyzed simultaneously.

very good, even slightly better for SH3-C. Interestingly the data reproduction obtained by analyzing the two proteins with a unique tensor is almost as good as the one obtained by the two separate analysis (total χ^2 of 136 for the separated analysis and 149 for the simultaneous one), again indicating that the relative orientation of the two proteins obtained from the RDC analysis is in agreement with the ^{15}N relaxation information and that the structure of the complex — the relative orientation of the two partners — can be

determined using the extrapolated ^{15}N relaxation rates, by determining separately the tensors and then aligning them. As the obtained tenors determined here present a clear axial symmetry, only the main axis can be easily determined leading to a potential indetermination of the orientation of the two protein against the rotation of one partner around the main axis. This degeneracy may be overcome by introducing supplementary restraints, e.g. ambiguous restraints derived from the chemical shift titrations.

In order to further investigate the orientational information presents in the R_2/R_1 ratio, a direct comparison with the available RDCs measured for the complex in PEG/hexanol mixture is presented in Figure 62. The

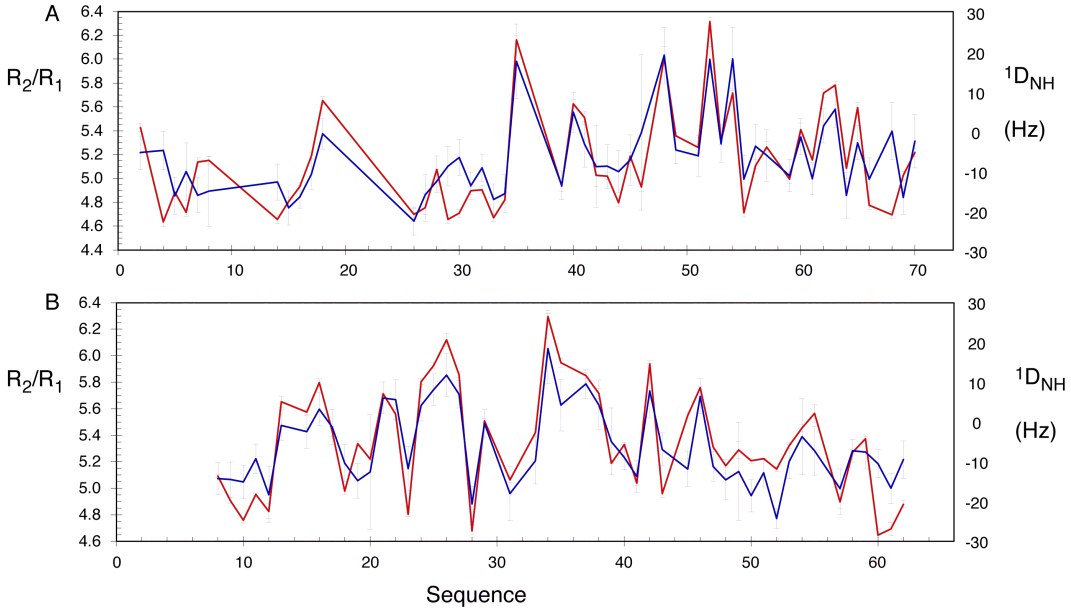

Figure 62 – Comparison of RDCs measured in PEG/hexanol (red) and R_2/R_1 ratios (blue) determined for the SH3-C-Ubiquitin complex: (A) Ubiquitin and (B) SH3-C

correspondence between these two completely independent data sets, neither of which are experimentally available, and both of which have been derived from multiple mixtures to access the values in the complex, is really quite striking. The amplitude of the two distributions cannot be compared, but the distribution profiles are extremely similar, as confirmed by the correlation plot presented in Figure 63. The information accessible by the two approaches is clearly of similar nature, but also complementary. The alignment tensor in a steric alignment media and the reorientation diffusion tensor has to present similarities, both being mainly determined by the shape of the considered system. Nevertheless those information are derived from completely different sources of information, by completely different experiments and both were extrapolated in the complex. Therefore the convergence of the two analyses bring important support to the two different

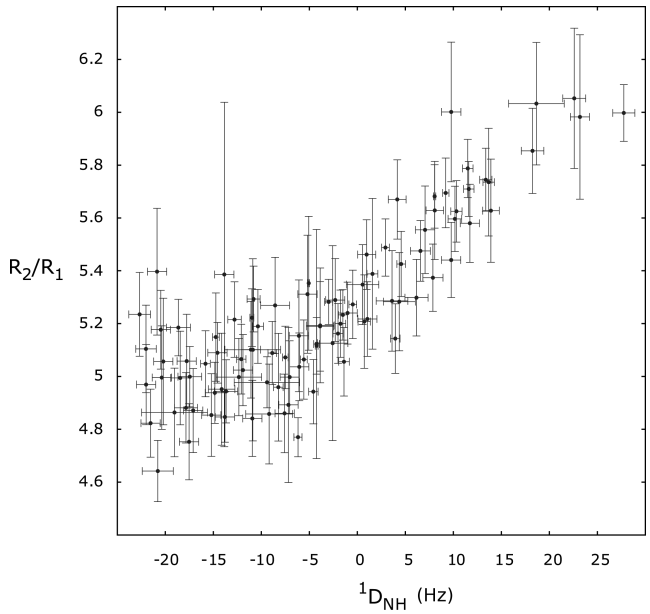

Figure 63 – Correlation of RDCs measured in PEG/hexanol and R_2/R_1 ratios determined for the SH3-C-Ubiquitin complex.

approaches. However the information content of the two sources of information — ^{15}N relaxation rates and RDCs — is by essence different and thus differences are expected. One of the most important sources of divergence between those two quantities is their different sensitivity to slow timescales dynamics. Nevertheless the structural information that modulates both ^{15}N relaxation rates and RDCs "dominate" the dynamical information, allowing this striking comparison. For this reason R_2/R_1 ratios have often been used similarly to RDCs in structural refinements of complexes.

8.3.9 *Determination of the Internal Dynamic of the Complex*

Knowing the properties of reorientation of the complex allows the determination of the local dynamics in the complex using a model-free approach. As the treatment of each amino-acid is independent, all residues are included: this allows the identification the excessive exchange contribution induced by the linear extrapolation. The results in terms of N_i-H_i^N order parameters $S^2_{NH,B}$ are shown in Figure 64 and compared to those of the free forms $S^2_{NH,F}$ (see Chapters 4 and 7).

The comparison of the N_i-H_i^N order parameters from the free and the bound forms leads essentially to similar results. A slight shift can be observed towards less dynamics in the complex, but often the differences are not large enough to be significant, considering experimental uncertainties. Some more important variations are localized in few regions of the two proteins, for Ubiquitin in the N-terminal part, in the 50-55 region and mainly in

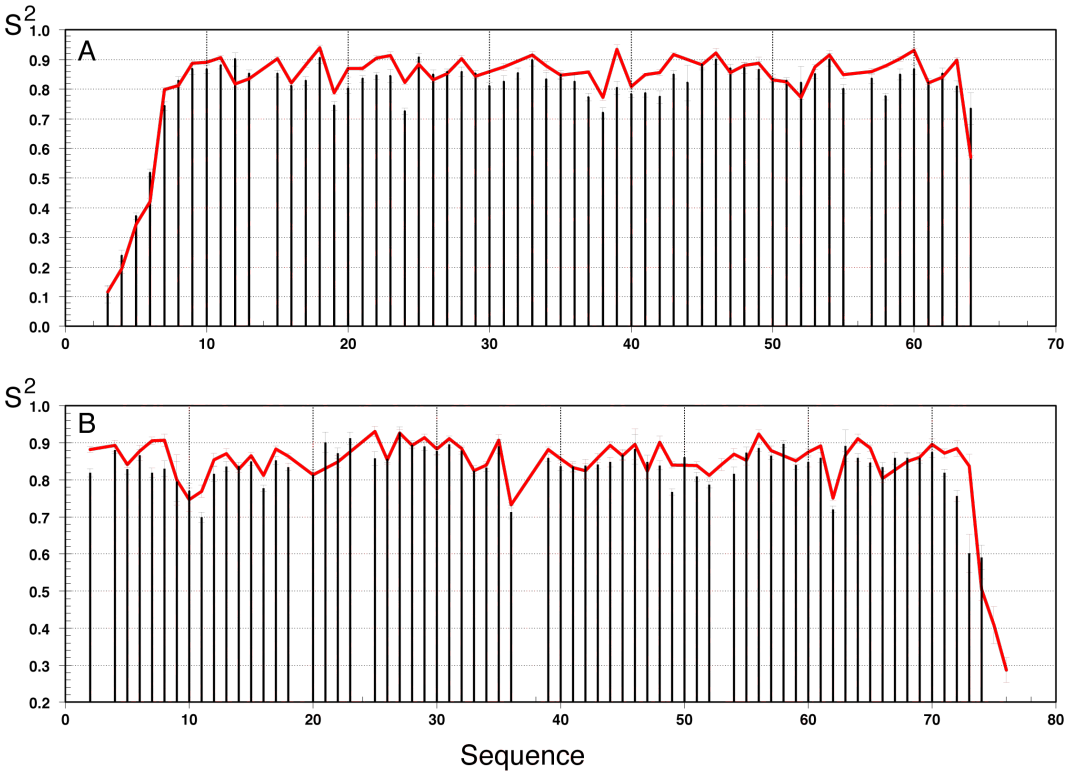

Figure 64 – Comparison of the ^{15}N relaxation derived order parameters of SH3-C and Ubiquitin in their free (black bars) and bound forms (red line): (A) SH3-C (B) Ubiquitin.

the C-terminal tail. For SH3-C they are clustered in the 18-25, 38-45 and 55-60 region. Those regions correspond mainly to those identified in the titration as being influenced by the complex formation. Those shifts may partially be due to a less accurate evaluation to the R_2 rates for those sites as they are all potentially influenced by the exchange due to the complex formation. Nevertheless if this bias may not be completely eliminated, it seems improbable that this effect appears systematically without inducing clear deviations for the data reproduction or for the comparison with RDCs and without generating exchange contributions in the model free analysis. A significant reduction of the dynamics is present in the C-terminal tail of Ubiquitin, which is known to play an important role in the complex formation [196]. A reduction of the dynamics at the interface of the complex is moreover physically reasonable: the appearance of interactions between the two partners reducing the mobility of the involved interfaces.

8.4 CONCLUSION

Physiologically important weak protein-protein interactions are remarkably common, for example for the dynamic rearrangement of multi-component

protein complexes involved in enzyme regulation and signal transduction. The low stability of weak complexes makes detailed structural study via X-ray crystallography almost impossible, and NMR spectroscopy is the technique of choice for studying low affinity interactions, providing atomic resolution characterization of molecular interfaces even for dissociation constants above 100 μM. Spin relaxation rates provide powerful constraints on intermolecular orientation in molecular complexes, but the unavoidable combination of free and bound contributions to the measured rate have until now precluded their use in the study of weak complexes.

In this chapter a weak complex between Ubiquitin and the SH3-C domain was investigated using ^{15}N relaxation rates. Using titration based approaches the R_1 relaxation rates and the R_2 rates for sites that do not exhibit exchange were accurately extrapolated in the complex. In favorable cases the exchange contribution due to the complex formation can be accurately determined allowing a precise determination of every R_2 rate in the complex and a precise estimation of the kinetic constants of the equilibrium. Here due to the properties of the considered complex (weakness of the interaction, specific rates of the formation-dissociation kinetics...) this determination could not be done with a sufficient accuracy and robustness. Thus alternative approaches were used by exploiting the information derived from the model-free analyses of the different mixtures and the complex. This allows a rough estimate of the kinetics of this system. Nevertheless an intriguing discrepancy between the chemical shifts obtained from the titration and those from the exchange contribution analysis remains and is still under investigation.

This analysis allows on to very accurately determine the rotational diffusion tensor of the complex. This analysis can be done using the two proteins separately and the obtained results are almost identical. The appearance of the anisotropy of the tensor can be followed from the free forms to the complex in the different mixtures. The obtained rates in the complex were compared to RDCs measured in steric alignments. The correlation observable by comparing the two datasets is striking, underlying the similar orientational information present in the two datasets, and demonstrating that the fluctuations measured in the R_2/R_1 ratios stem almost entirely from the orientation of the different relaxation mechanisms relative to the diffusion tensor.

Finally the approaches allowed the fast dynamics of the complex to be analyzed. The N_i-H_i^N order parameters are globally similar to those obtained in the free forms, except in the interacting regions, where a higher rigidity is observed in the complex. Nevertheless those changes are quite small and more significant differences might be visible at slower timescales. The

measurement of several RDC datasets in the complex, if experimentally challenging, would be of great interest to characterize the dynamics of this complex and would certainly bring further insights in the molecular recognition mechanisms.

Part IV

INTRINSICALLY DISORDERED PROTEINS AND ASTEROIDS

9

HIGHLY FLEXIBLE SYSTEMS: CHEMICALLY DENATURED AND INTRINSICALLY DISORDERED PROTEINS

ABSTRACT

Intrinsically disordered proteins are now recognized as playing important roles in the functioning of living organisms. Nevertheless these very flexible systems fall outside the classical paradigm of structural biology and require descriptions that can properly express their intrinsic conformational flexibility. Here some of the information that can be obtained using NMR spectroscopy is summarized and an ensemble based description, named Flexible-Meccano, for describing the conformational behavior of IDPs is presented.

9.1 INTRODUCTION

Intrinsically Disordered Proteins (IDPs) are now recognized to represent a significant fraction of all functional proteins. Even though this fraction is limited to a few percent in prokaryotic organisms, it is estimated to reach more than 30% for eukaryotic systems and even 40% of the human proteome [217]. The biological functions exerted by IDPs are very broad including cell cycle regulation, signaling, translation and transcription [217–219] and they play key roles in neurodegenerative diseases and cancer [217].

The inherent plasticity of IDPs allows them to sample more efficiently their surroundings and thereby increase the probability of interaction with one or several different biological partners [217], and they often fold upon binding [220, 221] although high flexibility can remain in the formed complexes [222]. IDPs have large interaction surfaces, without requiring excessively high molecular weights, and they are efficiently regulated through high turnover rates [217]. Due to their highly flexible nature, these proteins have until now escaped detailed biophysical characterization and it has

become clear that the very existence of this class of proteins necessitates a reassessment of the classical structure-function paradigm.

IDPs cannot be represented by a single, three-dimensional structure. Therefore, their characterization and biophysical properties have to be described using tools allowing for such flexible behavior. An ensemble description can be invoked, where the protein is assumed to interconvert between more or less equiprobable conformations. In general, descriptions based on probabilistic formalisms are expected to be more suitable for describing the properties of IDPs compared to standard structural approaches. Experimentally, many different techniques have been used to provide important insight into IDP biophysics [223] including Infra-Red (IR) Spectroscopy [224], Ciruclar Dichroïsm [225] and X-ray and neutron scattering [226, 227].

Considering IDPs as completely unfolded systems (i.e. as random-coil states) is an oversimplification as many IDPs contain significant amounts of residual structure [219, 225, 228]. This residual structure has been shown to be essential for controlling early molecular recognition events in IDPs undergoing disorder-to-order transitions upon binding to physiological partners. Therefore, many recent IDP developments have focused on characterizing such residual structure at atomic resolution [229, 230].

Thermally or chemically denatured proteins are also highly flexible and therefore can be expected to share some of the biophysical properties of IDPs. Some differences between IDPs and denatured proteins do however exist, e.g. the details of their conformational behaviour and their interactions with the solvent [223], however, the same techniques can be applied to both types of systems. The present chapter focuses on the use of NMR spectroscopy to characterize the dynamical behavior of IDPs and other unfolded states. Different kinds of information accessible by NMR will be presented and emphasis will be put on a recently developed ensemble description of the disordered state, the so-called Flexible-Meccano.

9.2 NMR AS A PROBE OF DENATURED AND INTRINSICALLY DISORDERED PROTEINS

NMR has emerged as the technique of choice for studying IDPs as it can provide both site-specific and long-range information in highly flexible systems by exploiting different kinds of interactions [229, 230]. This will be briefly explained in this section.

9.2.1 *Chemical Shifts*

The Chemical Shift (CS), which is the easiest measurable quantity in liquid state NMR, provides information about the local electronic environment of a studied nucleus [15, 231, 232]. In an unfolded system, a nucleus experiences different environments and the measured CS corresponds to the average over all conformations exchanging at timescales up to the millisecond.

The easiest way to extract information from CSs is to define the so-called random-coil values that correspond to the CSs of the nuclei in a completely unfolded chain. These random-coil values have to be obtained for each amino-acid and nucleus type and usually short peptides are used for this purpose [233]. Deviations from these values, the so-called Secondary Chemical Shifts (SCSs), can therefore be interpreted in terms of the existence of secondary structure propensities [233–235]. The use of several CSs per amino-acid, e.g. $^{13}C^{\alpha}$, $^{13}C^{\beta}$... allows a more precise estimation of the secondary structure propensity and limits the potential bias due to CS referencing errors [236, 237].

9.2.2 *J-Couplings*

The ^{3}J-coupling depends on the backbone ϕ-angles through an empirical Karplus-type relationship (see Chapter 5). As for chemical shifts, it is possible to estimate random-coil ^{3}J-coupling values and information about conformational propensities can be obtained through the deviations from these values [238, 239].

9.2.3 *Relaxation Measurements*

Spin relaxation measurements (e.g. ^{15}N R_1, R_2 and heteronuclear nOes) can be carried out in the unfolded state in order to obtain insight into the different motions occurring along the chain. Analysis of these relaxation rates is however complicated by the fact that the model-free approach is not directly applicable to unfolded systems as the internal and overall correlation functions cannot be clearly separated [230]. As an approximation, the overall correlation time can be modeled as a distribution of correlation times [240]. Information about slower motions at the µs-ms time scale has been obtained using relaxation-dispersion experiments [241] or R_2 measurements [242].

9.2.4 *Residual Dipolar Couplings*

As presented in Chapter 2, RDCs are sensitive probes of the structure and dynamics of a studied system. In principle this information exists both in folded and unfolded systems as long as RDCs can be measured. For completely unfolded proteins, the existence of non zero RDCs was at the beginning not obvious, as the flexibility of the peptide chain could potentially average the angular dependence of the RDCs to zero. However, it has become clear that RDCs in unfolded systems adopt bell-shaped curves with vanishing values at the extremities and non zero mean values in the center of the chain. This can be rationalized for a steric alignment medium by considering an internuclear vector, e.g. an N_i-H_i^N bond, in the center of the chain and one at the extremity. The alignment can be estimated by finding all possible bond vector orientations in the presence of the alignment medium and compare it to those in the absence of the medium[1]. In the absence of the medium the internal and external bond vectors have the same accessible space. Therefore, the vector with the most restricted available space when the medium is introduced will be more aligned and will exhibit larger RDCs values. The probability to find a conformation where the external amide N_i-H_i^N vector is close to the alignment medium elements (e.g. a bicelle) is much higher than for the internal vector and thus the central part of the chain should exhibit larger RDCs.

This qualitative reasoning has been more precisely studied by using a random walk model where the peptide chain is modeled as rigid segments [244–246]. This approach has provided an analytical expression of the bell-shaped curve that gives reasonable agreement with experimental data. A major drawback of this approach is that each amino-acid is treated in the same way and therefore does not allow different conformational propensities for each amino-acid. In order to obtain a quantitative description of the unfolded state, it is essential to take into account amino-acid specific conformational propensities.

The deviation from the completely unfolded state can be monitored as modulations of this bell-shaped curve. For example, studies of the denatured states of apo-myoglobin [247] and acyl-CoA binding protein [248] reveal deviations of the $^1D_{NH}$ couplings from their negative bell-shaped profile. In some cases the $^1D_{NH}$ values even become positive showing that the proteins contain significant amounts of residual structure in the denatured state. The observed change in sign can be explained as follows. As shown in Chapter 2, a RDC can be expressed as $\left\langle d_{NH} P_2(\cos\theta_{NH})\right\rangle$, where θ_{NH} is the angle between the N_i-H_i^N vector and the magnetic field $\mathbf{B}_0$ and P_2 is the

1 This is the principle underlying the PALES approach [243].

second order Legendre polynomial. For a N_i-H_i^N internuclear vector, d_{NH} is positive so the sign of the RDC will be given by $\langle P_2(\cos\theta_{NH})\rangle$. If we consider a steric alignment medium where an elongated molecule aligns in the $\mathbf{B}_0$ field direction, as shown in Figure 65, the N_i-H_i^N internuclear vector will mainly be orthogonal to the $\mathbf{B}_0$ field ($\theta_{NH} \simeq 90\,°$) and the corresponding $^1D_{NH}$ will be negative. In the case of an elongated chain with α-helical or turn-like structure the N_i-H_i^N internuclear vector will mainly be parallel to the $\mathbf{B}_0$ field ($\theta_{NH} \simeq 0\,°$) leading to positive RDCs.

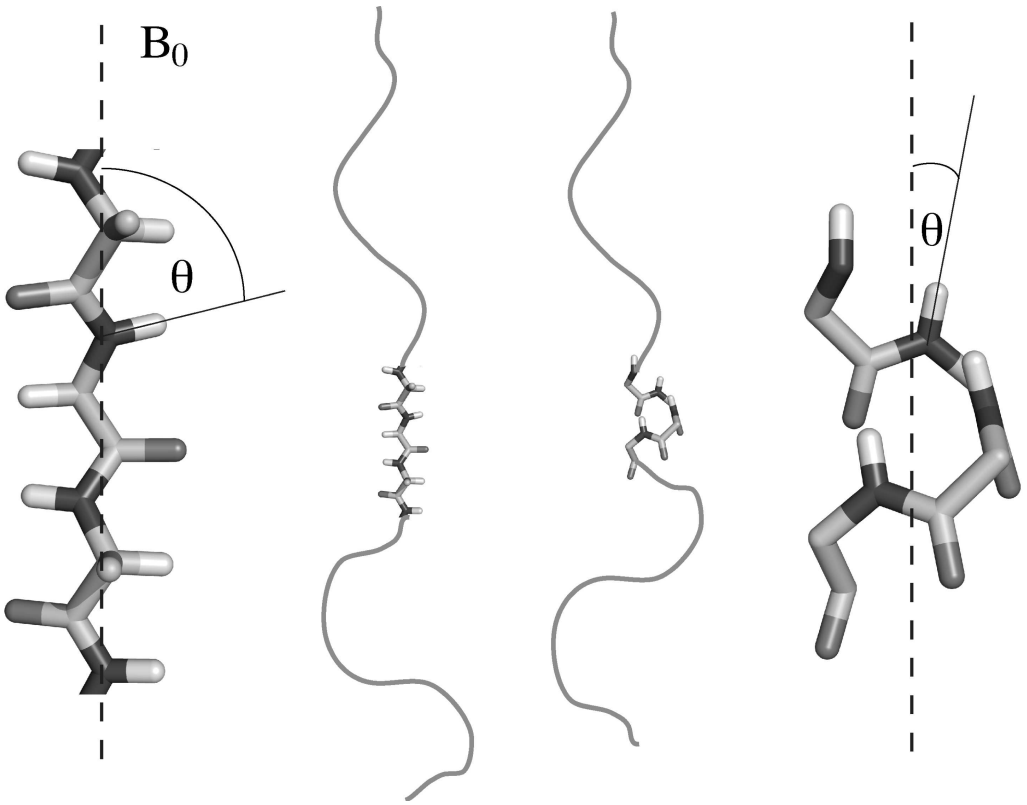

Figure 65 – Figurative representation of the effect of the absence (A) or presence (B) of an α-helical motif in an elongated structure on the orientation of N_i-H_i^N amide internuclear vector. The alignment direction is supposed to be parallel to the $\mathbf{B}_0$ field.

It is worth noting that RDCs for unfolded proteins are often measured in steric alignment media because tensor estimation is relatively simple relying on the over-all shape of the molecules [243]. Tensor estimation is more complicated in electrostatic alignment media as the distribution of charges in the protein has to be taken into account [249]. Furthermore, as IDPs often exhibit a high proportion of charged residues [217], strong interactions with electrostatic alignment media may be more difficult to avoid.

9.2.5 *Paramagnetic Relaxation Enhancement*

A convenient way to extract long-range distance information in the unfolded state is to use PREs induced by an artificially incorporated unpaired electron (see Section 1.3). The presence of an unpaired electron enhances the relaxation rates of nuclei located up to 30 Å from the electron. The relaxation enhancement of nuclei located close to the unpaired electron can be so significant that the NMR signals become broadened beyond detection. Thus, PREs are considered to bring relevant distance information in the range 5-30 Å from the unpaired electron [250].

Classical approaches to obtain such information is to attach a paramagnetic tag on specific places in the poly-peptide chain. A commonly used paramagnetic probe is the MTSL (1-oxyl-2,2,5,5-tetramethyl-Δ3-pyrroline-3-methyl)-methanethiosulfonate) spin label that can be attached to Cysteine sulfur atoms. Native Cysteines or Cysteines that have been introduced into the protein by site-directed mutagenesis can be used as MTSL anchor sites [250–252]. Using multiple anchor sites enhances the available distance information and possibly allows an accurate characterization of the distance distributions within the protein. Nevertheless the use of mutagenesis and the introduction of a tag can modify the biophysical properties of the protein and caution has to be taken to minimize potential perturbations.

In general, PREs are obtained by measuring the peak intensity ratio between two different spectra, one in presence of the paramagnetic center, one in the absence of this center [250, 252] or by measuring directly R_2 relaxation rates [250, 253]. For the MTSL spin label, the oxidized form of the nitroxyl group is paramagnetic, whereas the reduced form is diamagnetic. Thus, the PREs can be measured using the same protein sample by reducing or oxidizing the paramagnetic center.

Pioneering work has been achieved by Gillespie and Shortle [254], with the analysis of denatured staphylococcal nuclease, where the authors identified deviations from the completely unfolded state by comparing the electron-nucleus correlation time profile obtained experimentally and the one predicted for a completely random walk model. The theory underlying their observation was later further investigated with different analytical models and molecular simulations [255]. Further work characterizes distance information in unfolded states by incorporating PREs as distance restraints in restrained molecular dynamics simulations (see Section 3.4) [252, 256, 257] or by interpreting them in terms of probability distribution functions [258]

9.3 FLEXIBLE-MECCANO

9.3.1 *Principle*

FLEXIBLE-MECCANO is an ensemble based description of the unfolded state [103] that consists of generating a large ensemble of structures whose average properties represent the unfolded state.

This method is derived from the MECCANO algorithm [165] (see Section 3.6.3), that allows the construction of a protein structure by sequentially orienting one peptide plane after another. In the MECCANO approach, applied to folded proteins, the plane is oriented to be in agreement with measured RDCs, whereas in the FLEXIBLE-MECCANO approach no similar restraints are used. The planes have an ideal geometry and are oriented by following two rules:

1. The backbone (ϕ, ψ)-angles are selected in a database that is assumed to describe the unfolded state. This database is amino-acid specific, with some additional special cases such as amino-acids preceding proline are treated as specific amino-acids due to their restricted sampling. This database is an important feature of the model, and the first database was constructed by including all (ϕ, ψ)-combinations adopted by residues in loop regions of 500 high resolution (less than 1.8 Å) X-ray structures.

2. A conformation, i.e. a (ϕ, ψ)-combination, can be selected only if it does not generate steric clashes. Clashes are estimated by using a residue specific exclusion volume modeled as a sphere centered on the C^{β} atom (C^{α} atom for glycines). The radius of the spheres are amino-acids specific and their values have been fixed according to the literature [259].

For each conformer the desired quantities, e.g. ${}^{1}D_{NH}$ couplings, are calculated and averaged over all the members of the ensemble.

Concerning RDCs, they are estimated using a static description (see equation 2.30), with the conformational disorder of the unfolded protein represented by the ensemble of states, and the tensor properties are determined by either PALES [243], an approach that estimate steric alignment tensors on the basis of the shape of the protein, or by using the similarity between alignment and the gyration tensor [260].

9.3.2 *Applications*

The FLEXIBLE-MECCANO approach has been successfully applied to various systems. The first application was the nucleocapsid-binding domain of Sendai virus phosphoprotein P_X. Its biological partner N_{TAIL}, the C-terminal domain of the Sendai virus nucleoprotein, has also been investigated using FLEXIBLE-MECCANO based approaches [261, 262].

The protein α-Synuclein, was also investigated [263], where long-range contacts were necessary to reproduce experimental RDCs and during the study of the Tau protein [264], the combination of FLEXIBLE-MECCANO and molecular dynamics simulations allowed the identification of β-turns in the vicinity of the functionally important microtubule binding, and aggregation nucleation sites.

It was also applied to urea-denatured Ubiquitin [265], where multiple RDCs measured for each peptide unit clearly indicated that the statistical coil model of the unfolded state needed to be refined to take into account the more extended nature of backbone conformational sampling in the presence of urea.

9.4 CONCLUSION

IDPs are challenging biophysical systems as their intrinsic flexibility requires that biophysical approaches that have been developed to study folded systems are not longer appropriate and need to be significantly adapted.

Their biological roles can be complex and the associated molecular processes are often completely unknown. As they fall outside of the classical structural biology paradigm, a large set of question appears, concerning their biomolecular behavior or even concerning the adequate way of studying such systems. Nowadays the understanding of the biophysical properties of the unfolded state is not sufficiently established. The use of approaches based only on classical biology hypothesis could miss some still unknown features of the unfolded state.

Furthermore, the philosophy of IDP studies seems to have some symmetry with studies of the dynamics of folded proteins. Two extreme situations exists: the perfectly static description for folded systems and the random-coil model for unfolded systems. In both case deviations are observed and are biologically important. For folded systems the introduction of conformational dynamical disorder is necessary to obtain an accurate picture of their biophysical properties. For the unfolded state the random-coil has

appeared as an oversimplified description and precise descriptions are required to identify the presence of functionally important residual order.

Therefore the two following chapters will deal with the presence of conformational order in denatured and unfolded systems, characterized with a non hypothesis driven approach: the first will focus on local characterization and the second on long-range manifestation of this conformational (dis)order.

10

CHARACTERIZATION OF LOCAL ORDER IN UNFOLDED SYSTEMS

ABSTRACT

The existence of local conformational propensities in unfolded system may be extremely important for their biological function. Characterizing this residual structure form experimental data remains difficult without using hypothesis driven approaches. Here a novel method is developed to obtain local conformational sampling of unfolded proteins. As the studied systems present an extremely large number of degree of freedom the approach is extensively tested in order to determine the extent to which the reproduction of experimental data ensures a meaningful representation of the conformational distribution. The approach is first developed for urea-denatured Ubiquitin by analysing an extensive set of RDCs and then used to describe N_{TAIL} conformational sampling on the basis of heteronuclear chemical shifts.

10.1 INTRODUCTION

The FLEXIBLE-MECCANO approach presented in the previous chapter (see Section 9.3) has led to a significant improvement in our understanding of the unfolded state. Thus, if we assume that the FLEXIBLE-MECCANO database is reasonable, the ensemble constructed by FLEXIBLE-MECCANO provides a reliable representation of the disordered state and includes all conformations exchanging at timescales up to the millisecond.

Nevertheless, this approach alone is unable to rationalize deviations from the random-coil state. Previous studies have mainly used a rational, hypothesis-based approach to the interpretation of RDCs in unfolded states, calculating explicit ensembles containing tens of thousands of conformers from different conformational sampling regimes and comparing the ensemble-averaged couplings to experimental data [264, 265]. In this study we are interested in

investigating the possibility of defining the conformational sampling of the peptide chain directly from the experimental NMR data at amino-acid resolution. In order to do this we develop a novel algorithm, called ASTEROIDS (A Selection Tool for Ensemble Representations Of Intrinsically Disordered State), to select from a large pool of possible conformers, created using the algorithm FLEXIBLE-MECCANO, the sub-ensemble that best describes the data.

First, we test this approach to determine how well the fitting of RDCs to reduced conformational ensembles containing few conformers can correctly reproduce the peptide backbone conformational behavior of the protein. Having established approaches that allow accurate mapping of conformational space, the protocol is applied to the analysis of urea-denatured Ubiquitin using RDCs and with the analysis of the CSs of N_{TAIL}, the C-terminal domain of Sendai virus nucleoprotein (see Annexe B).

This chapter summarizes two pieces of work in which I was strongly implicated, but for which I was not the only principle author. Major part of the development of ASTEROIDS approach was made by GABRIELLE NODET and this application to Chemical Shifts was leaded by MALENE RINGKJØBING JENSEN.

10.2 MATERIALS AND METHODS FOR UREA-DENATURED UBIQUITIN ANALYSIS

10.2.1 *Experimental Data*

Experimental RDCs of urea-denatured Ubiquitin were all measured by Meier et al. [265]. The data set is made up of ${}^{1}D_{NH}$, ${}^{1}D_{C'C^{\alpha}}$, ${}^{1}D_{C^{\alpha}H^{\alpha}}$, ${}^{3}D_{H^{N}H^{\alpha}}$, ${}^{4}D_{H_i^N H_{i-1}^{\alpha}}$, ${}^{5}D_{H_i^N H_{i+1}^N}$, and ${}^{8}D_{H_i^N H_{i+2}^N}$, where all couplings were measured by aligning the protein in stretched polyacrylamide gel at pH = 2.5, 8 M urea.

10.2.2 FLEXIBLE-MECCANO *Ensemble Generation and RDCs Calculations*

FLEXIBLE-MECCANO calculations were carried out as explained in Section 9.3.

For RDCs, 12 000 urea-denatured Ubiquitin structures were generated, half of them using the standard random-coil database, half of them using a more extended database where the region with $50° < \phi < 180°$ in Ramachandran space is more populated (78% instead of 59%). The extended database has

been shown to be more appropriate for describing urea-denatured proteins compared to the standard random-coil database [265].

RDCs were calculated using PALES [243]. Two methods were used to simulate RDCs. One where the tensor of the protein was estimated using the whole molecule (standard PALES protocol) and one where the tensor was determined using a Local Alignment Window (LAW) [266]. For each amino-acid i in the protein, a local alignment tensor was estimated using short windows of 3, 9, 15 or 25 amino-acids in length centered on i and RDCs were calculated for this amino-acid i only. This protocol was repeated for all amino-acids in the peptide chain. Dummy Alanines were added to the C- and N-terminal ends of the protein in order to be able to use this procedure for the whole peptide chain. RDCs were then averaged over all ensemble members. For the LAW based protocol, the obtained RDCs were multiplied by an appropriate hyperbolic cosine baseline, defined for each amino-acid i as:

$$B(i) = 2b\cosh(a(i - m_0)) - c \tag{10.1}$$

with m_0 corresponds to the middle of the sequence and a, b, c are adjustable parameters that are optimized for the different types of couplings. Thus, the i-th coupling between spins I and S can be expressed in the LAW based protocol as:

$$D_{IS}(i) = |B(i)|D_{IS}^{LAW} \tag{10.2}$$

10.2.3 *Simulated Data*

Two synthetic datasets of RDCs were simulated using the structures of the standard sampling or the structures of the extended sampling generated using FLEXIBLE-MECCANO (see Section 10.2.2). ${}^{1}D_{NH}$, ${}^{1}D_{C^{\alpha}H^{\alpha}}$, ${}^{1}D_{C'C^{\alpha}}$, ${}^{3}D_{H^{N}H^{\alpha}}$, ${}^{4}D_{H_i^{N}H_{i-1}^{\alpha}}$, ${}^{5}D_{H_i^{N}H_{i+1}^{N}}$ and ${}^{8}D_{H_i^{N}H_{i+2}^{N}}$ were simulated for all the structures and the synthetic datasets correspond to the average over 5000 structures.

10.2.4 ASTEROIDS

ASTEROIDS is an approach based on a genetic algorithm that allows selection of a sub-ensemble from a large pool of structures that on average agrees with available experimental data for example from NMR.

Genetic algorithms, which are very popular evolutionary algorithms, were developed by Holland and co-workers [267]. The idea of such an algorithm

is to mimic biological evolution [268] to select from a large population the sub-ensemble which is the best adapted to a given problem.

The aim of ASTEROIDS is to select an ensemble of N conformers within a large pool of structures obtained using FLEXIBLE-MECCANO.

Initially, a population of P solutions (called individuals) is generated at random, where each individual is made up of an ensemble of N structures. The aim of the algorithm is to select the individual that is best suited for representing the experimental data. Thus, the algorithm generates a population of solutions that evolve by iteratively repeating a step of population increase and a step of population decrease.

In order to increase the population of solutions, three evolution operators are used:

1. Random generation of solutions: as in the first step P individuals are randomly generated by selecting structures within the large pool of structures initially obtained by FLEXIBLE-MECCANO.

2. Reproduction or crossing: the solutions of the previous step are randomly paired and called parents. P child solutions are obtained by selecting N structures in the pool of structures obtained from the two parents.

3. Mutation: each individual is mutated by randomly changing 1% of the structures (or at least one structure). This mutation can be done by substituting structures with some coming from the pool consisting of all the structures present in the population of the previous step and is called an internal mutation. Otherwise mutation can be done by using structures that are not present in the previous population, and is called an external mutation. P individuals are obtained by both mutation protocols.

During the evolution, it is forbidden to obtain two identical individuals and each individual cannot contain the same structure twice.

After the evolution step, the population contains 5 P individuals. This population is then randomly subjected to T tournaments. In each tournament individuals are sorted according to a fitness function which is a classical χ^2:

$$\chi^2 = \sum_i \left(\frac{X_i^{calc} - X_i^{exp}}{\delta_i} \right)^2 \tag{10.3}$$

where i run over all the experimental measurements, e.g. RDCs or CSs.

The number of individuals surviving the selection according to the fitness function is controlled by the number of tournaments in order to reach in total P solutions. By varying the number of tournaments the selection pressure can be adjusted. Therefore, to avoid premature convergence in a local minimum, the number of tournaments is decreased during the evolution, with a typical variation from 100 to 50, 25, 20, 10, 2 and finally 1.

The evolution is repeated several times (typically a few thousand) in order to ensure convergence of the results.

For the RDCs analysis, the weights of the couplings were fixed to 1.0 for ${}^{1}D_{NH}$ and ${}^{4}D_{H_i^N H_{i-1}^\alpha}$, 2.0 for ${}^{1}D_{C^\alpha H^\alpha}$, 0.5 for ${}^{1}D_{C'C^\alpha}$, ${}^{3}D_{H^N H^\alpha}$ and ${}^{5}D_{H_i^N H_{i+1}^N}$ and 0.33 for ${}^{8}D_{H_i^N H_{i+2}^N}$ and the number of individuals per generation P to 100. The RDCs were calculated by linear averaging and a scaling factor was applied to obtain agreement with the experimental RDCs range, i.e. the level of alignment. Here, two scaling factors were used: K_1 for ${}^{1}D_{NH}$, ${}^{1}D_{C^\alpha H^\alpha}$, ${}^{1}D_{C'C^\alpha}$ couplings and K_2 for ${}^{3}D_{H^N H^\alpha}$, ${}^{4}D_{H_i^N H_{i-1}^\alpha}$, ${}^{5}D_{H_i^N H_{i+1}^N}$ and ${}^{8}D_{H_i^N H_{i+2}^N}$ couplings. Convergence was achieved after 2000 iterations.

10.2.5 *Ramachandran Partition*

In order to describe the conformational sampling of the different ensembles, the Ramachandran space is divided into four quadrants, as shown in Figure 66, defined as:

- α_L: $\phi > 0^\circ$
- α_R: $\phi < 0^\circ$ and $-120^\circ < \psi < 50^\circ$
- β_P: $-90^\circ < \phi < 0^\circ$ and $\psi < -120^\circ$ or $\psi > 50^\circ$
- β_S: $-180^\circ < \phi < -90^\circ$ and $\psi < -120^\circ$ or $\psi > 50^\circ$

For each quadrant q a population p_q, for the amino-acid j can be defined. For estimating the capability of an ensemble to reproduce conformational sampling of a simulated target, a χ^2_{Ram} can be introduced as:

$$\chi^2_{Ram} = \sum_{j,q} \left(p^{calc}_{q,j} - p^{ref}_{q,j} \right)^2 \tag{10.4}$$

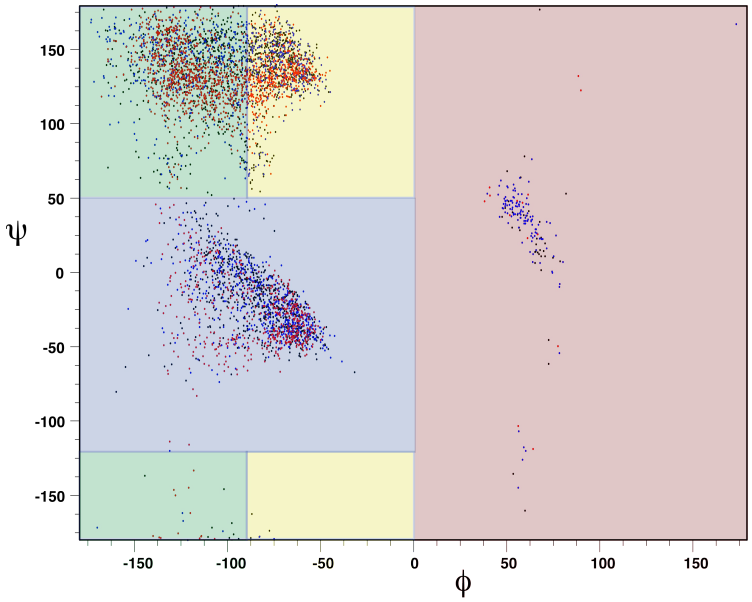

Figure 66 – Definition of the four Ramachandran space quadrants. Regions are α_L (red), α_R (blue), β_P (yellow) and β_s (green). Dots represent standard statistical coil distributions of Valine (red), Lysine (blue) and Leucine (black).

10.2.6 *Radius of Gyration Calculation*

The radius of gyration R_g of a molecule can be obtained using [269]:

$$R_g = \sqrt{\frac{\sum_i m_i(\mathbf{r_i} - \mathbf{r_c})^2}{M}} \tag{10.5}$$

where $\mathbf{r_i}$ is the position and m_i is the mass of nuclei i of the molecule, $M = \sum_i m_i$ is the total mass of the molecule and $\mathbf{r_c}$ is the center of mass of the molecule defined as:

$$\mathbf{r_c} = \frac{1}{M}\sum_i m_i \mathbf{r_i} \tag{10.6}$$

10.3 RESULTS AND DISCUSSION FOR UREA-DENATURED UBIQUITIN RDC ANALYSIS

As we have seen in the previous chapter, RDCs can be simulated from explicit molecular ensembles of disordered proteins using shape-based considerations of the alignment properties of each copy of the molecule, and the average couplings can be predicted by taking the mean over the entire ensemble. Comparison of such predictions with experimental data has revealed the unique sensitivity of RDCs to local and global sampling properties of highly disordered proteins. A key disadvantage of this approach is the number of structures that need to be treated, before the average RDC

value converges to a non-fluctuating value. This number can reach many tens of thousands in proteins of 100 amino-acids. It has recently been proposed that convergence of RDCs towards experimental data can be achieved with a smaller number of conformers if the protein is divided into short, uncoupled segments [265] — LAWs — and the RDCs are calculated using the alignment tensor of these segments (see Section 10.2.2).

The ability to describe the conformational properties with ensembles containing fewer structures will of course make any ensemble selection procedure more tractable, and is therefore an attractive prospect. In general however RDCs are affected both by the local conformational sampling, and the chain-like nature of the unfolded protein, that induce an effective baseline. Long-range information is necessarily absent from an approach that only employs LAWs to predict the RDCs. If this approach is employed the simulated data need to be corrected for the effects of the unfolded chain.

10.3.1 *Separation of Local and Global Effect on RDCs in Unfolded Proteins*

As previously discussed, RDCs in IDPs have a bell shaped profile. We have simulated ensemble-averaged RDCs for poly-Valine chains of differing lengths. This profile can for a given unfolded system be relatively well parametrized by an hyperbolic cosine curve (see equation 10.1). This parameterization can be seen in Figure 67 for a poly-Valine chain of 76 amino-acids in length. We have therefore investigated whether the RDC profile for a given unfolded protein can be separated into contributions form the local environment and from the global contribution due to the unfolded chain.

The use of this baseline takes into account the effect of the position of the amino-acid in the chain i.e. the contribution of the global "order" present at a given position in the chain. The local contribution can then be described using the local alignment window (LAW) approach (see Section 10.2.2). This approach estimates the alignment tensor for small segments of the peptide chain instead of using the whole molecule. This corresponds to defining a persistence length beyond which the neighboring residues have no influence on the conformational sampling.

The advantage of decoupling the two effects (local and global) is that it allows a description of the system using a smaller number of structures. This can be seen from Figure 68A, where the convergence of $^1D_{NH}$ couplings are presented for different LAWs. This acceleration of convergence is mainly due to the decrease of the RDC range as the LAWs get smaller, because the alignment tensor eigenvalues decrease with decreasing chain length (see Figure 68B).

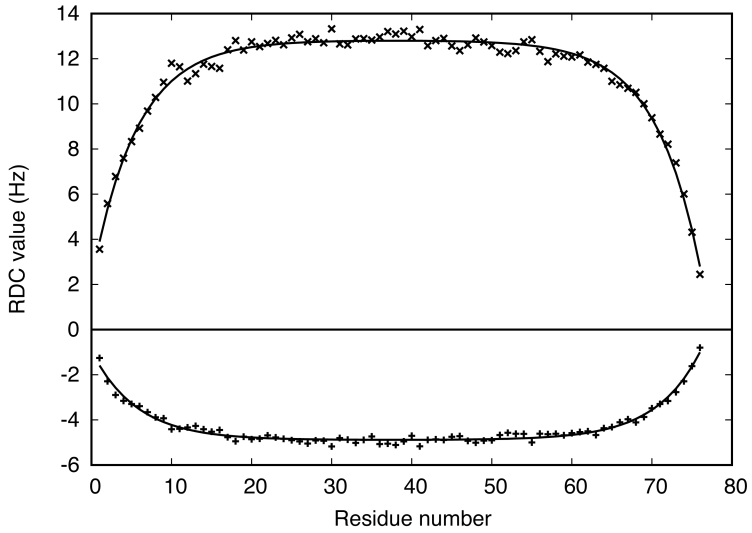

Figure 67 – Parameterization of the RDC bell-shaped curve in unfolded proteins. Crosses indicate ${}^1D_{C'C^\alpha}$ (positives values) and ${}^1D_{NH}$ (negatives values) simulated with 100 000 FLEXIBLE-MECCANO structures of a poly-Valine chain of 76 amino-acids in length. Solid line indicates the parameterization obtained using a hyperbolic cosine description.

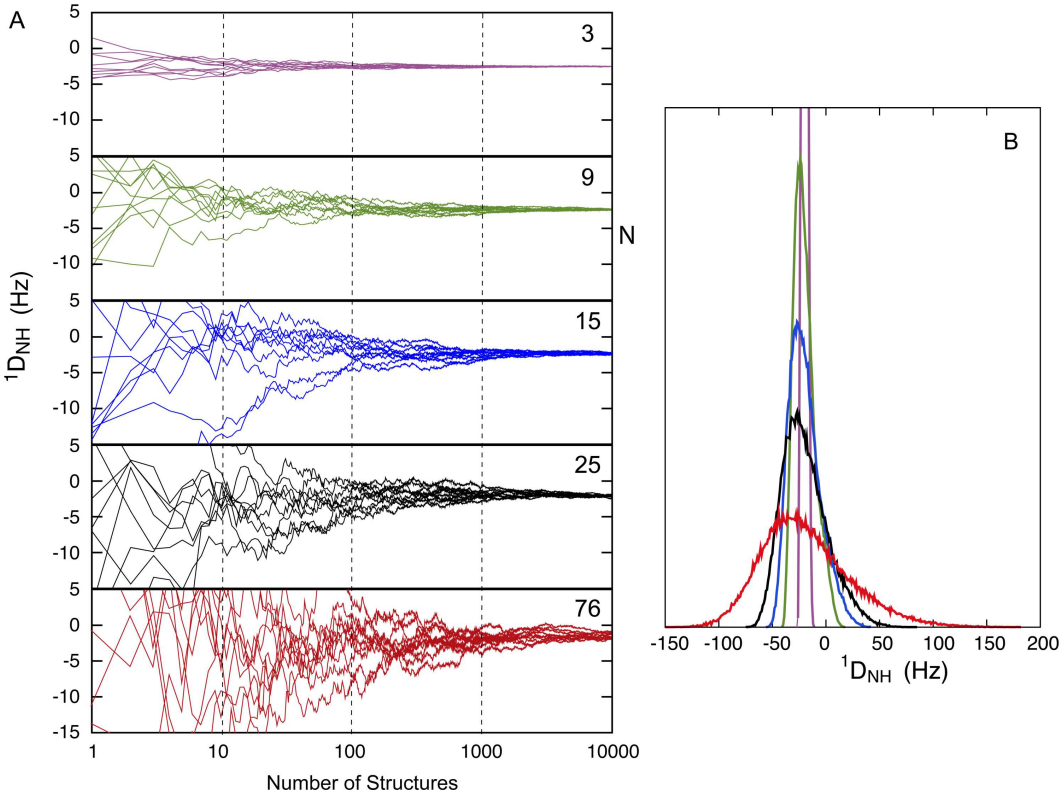

Figure 68 – Convergence of ${}^1D_{NH}$ couplings for different LAWs of length 3 (purple), 9 (green), 15 (blue), 25 (back) residues or from the full length protein (red) using a global alignment tensor. (A) Averaged values obtained for increasing number of structures for ten different simulations of ${}^1D_{NH}$ for the amino-acid 41 of Ubiquitin. (B) Distribution of RDCs obtained from the same simulations.

The length of the LAW has to be set to a reasonable value that allows to properly take into account neighboring effects. If the LAWs become too short, the RDC distribution cannot be correctly reproduced (data not shown). A value for the length of the LAW that offers both efficiency and accuracy was found to be around 15 amino-acids. Comparison of RDCs obtained using 15 amino-acid LAWs and obtained using a global alignment tensor is presented in Figure 69. It is seen that the separation of local and global effects is reasonable and provides similar accuracy of the predicted RDCs as the simulation using a global alignment tensor.

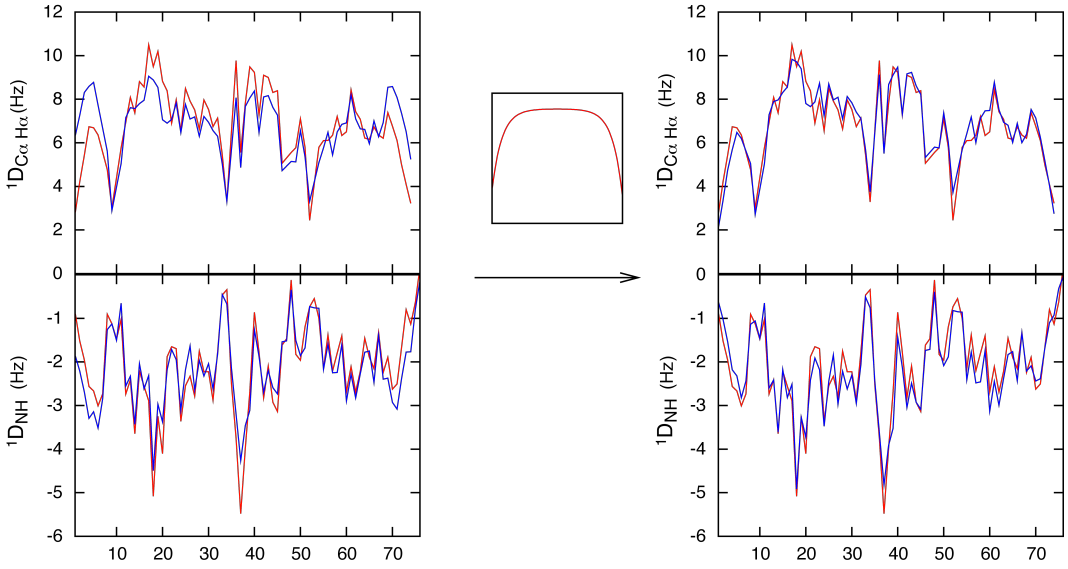

Figure 69 – Effect of separating local and long-range effect on RDCs. RDCs simulated with a full length approach (red) or with the LAW (blue). The LAW approach is used with (right) or without (left) baseline multiplication.

10.3.2 *Testing* ASTEROIDS *on Simulated Data*

In order to test the capability of the ASTEROIDS ensemble selection procedure to reproduce RDCs and conformational sampling, the approach was tested on the two simulated datasets. The obtained results are of similar quality and therefore only the results with the standard sampling (see Section 10.2.2), will be presented.

The number of structures used in the ASTEROIDS selected ensemble is a key parameter in the analysis that is difficult to estimate *a priori*. Thus, the influence of this parameter was investigated. Results in terms of RDCs and conformational sampling reproduction can be seen in Figure 70. χ^2_{RDC} relates to the reproduction of the data while χ^2_{Ram} reports on the validity of the resulting sampling, that is how close it is to the conformational sampling used to predict the data. Both χ^2_{RDC} and χ^2_{Ram} show a rapid decrease for

small ensemble sizes. However the slope of the drop is clearly less steep for χ^2_{Ram}, indicating that it takes until something in the range of 200 structures before the conformational sampling is reproduced accurately. We note that this number is an order of magnitude higher than some ensemble sizes that have been proposed in similar ensemble selection procedures. More or less identical results are obtained for the LAWs of 9 and 15 amino-acids in length, however, the simulation where a global alignment tensor is used the reproduction of both RDCs and conformational sampling is clearly worse.

The last point indicates that the estimation of the LAWs for full ensemble description remains valid for ASTEROIDS selection procedure. On the contrary selection of tractable size with the full length approach does not lead to satisfying results.

These result show that averaging regimes exist where the data can be reproduced within experimental error, but the conformational sampling regime is not accurately reproduced. This is a problem of solution uniqueness in the resolution of ill-posed problems and that indicates the clear necessity of performing extensive simulation before applying such approaches to the description of systems with far more degrees of conformational freedom than available data.

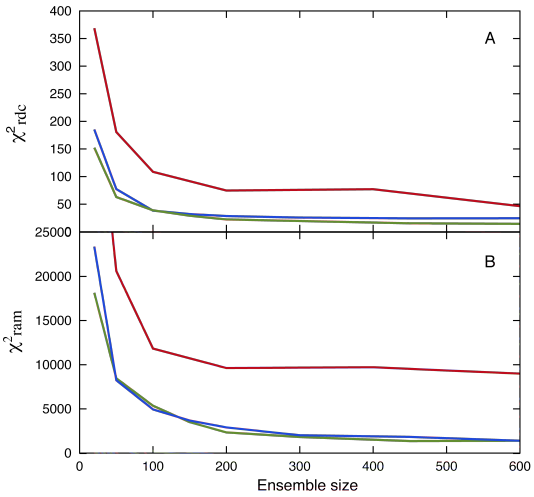

Figure 70 – RDC χ^2_{RDC} (upper panel) and conformational sampling χ^2_{Ram} (lower panel) reproduction of simulated data using the ASTEROIDS approach as a function of the number of structures in the ensemble. Two LAWs are presented 9 (green) and 15 (blue) amino-acids and full-length (red).

The size of the ensemble retained was therefore 200 structures and corresponding results in terms of RDCs and conformational sampling are presented in Figure 71 and compared to the results obtained for an ensemble of 20 structures. The accurate data reproduction and conformational sampling demonstrate the capacity of ASTEROIDS to select representative

ensembles of unfolded states. Although the fit is significantly poorer in the case of 20 structures, the overall features are actually quite well reproduced, and the quality of the fit would probably be considered acceptable in the presence of commonly encountered levels of experimental noise. However the conformational sampling is, in this case, very poorly reproduced.

10.3.3 *Applying* ASTEROIDS *on urea-denatured Ubiquitin Data*

The ASTEROIDS approach was then applied to the experimental RDC dataset of urea-denatured Ubiquitin. The pool of structures was made up of the 12,000 FLEXIBLE-MECCANO structures (see Section 10.2.2) and an ensemble of 200 structures was selected using ASTEROIDS. In order to properly reproduce the data, two scaling factors were used: $K_1 = 0.58$ for ${}^1D_{NH}$, ${}^1D_{C^\alpha H^\alpha}$, ${}^1D_{C'C^\alpha}$ couplings and $K_2 = 0.96$ for ${}^3D_{H^N H^\alpha}$, ${}^4D_{H_i^N H_{i-1}^\alpha}$, ${}^5D_{H_i^N H_{i+1}^N}$ and ${}^8D_{H_i^N H_{i+2}^N}$ couplings. Results are shown in Figure 72.

Good RDC reproduction is obtained when using the two different scaling factors. The necessity of differentiating covalently bound and inter-protons RDCs may be due to the presence of dynamics not explicitly taken into account in the applied model. This dynamics could influence covalently bound and inter-proton RDCs differently.

Robustness and accuracy of the approach were tested using Monte-Carlo simulations and cross-validations. The cross-validations (see Figure 73A) indicate that correct data reproduction can be obtained for unused couplings (representing 10% of the data) and that no RDC type is more important than the other to define urea-denatured Ubiquitin conformational sampling (data not shown). In addition, the cross-validations confirm that 200 structures is enough to reach a plateau value of the indirect χ^2_{RDC} (see Figure 73B). Monte-Carlo simulations (data not shown) were used to estimate the uncertainty (around 3%) of the population in each quadrant in the Ramachandran space.

Concerning conformational sampling, the obtained ensemble exhibits differences from the random-coil distribution that are larger than the estimated uncertainty (see Figure 72). A clear tendency to sample more in the β_P and β_S regions and less in the α_R region is observed. This corresponds to having more extended conformations. Previous studies [265] already introduced this hypothesis to explain the same experimental dataset, but here the results were obtained by a direct analysis of the data.

A site specific description of the conformational sampling is presented in Figure 74. It is seen that threonine, glutamic acid and arginine often

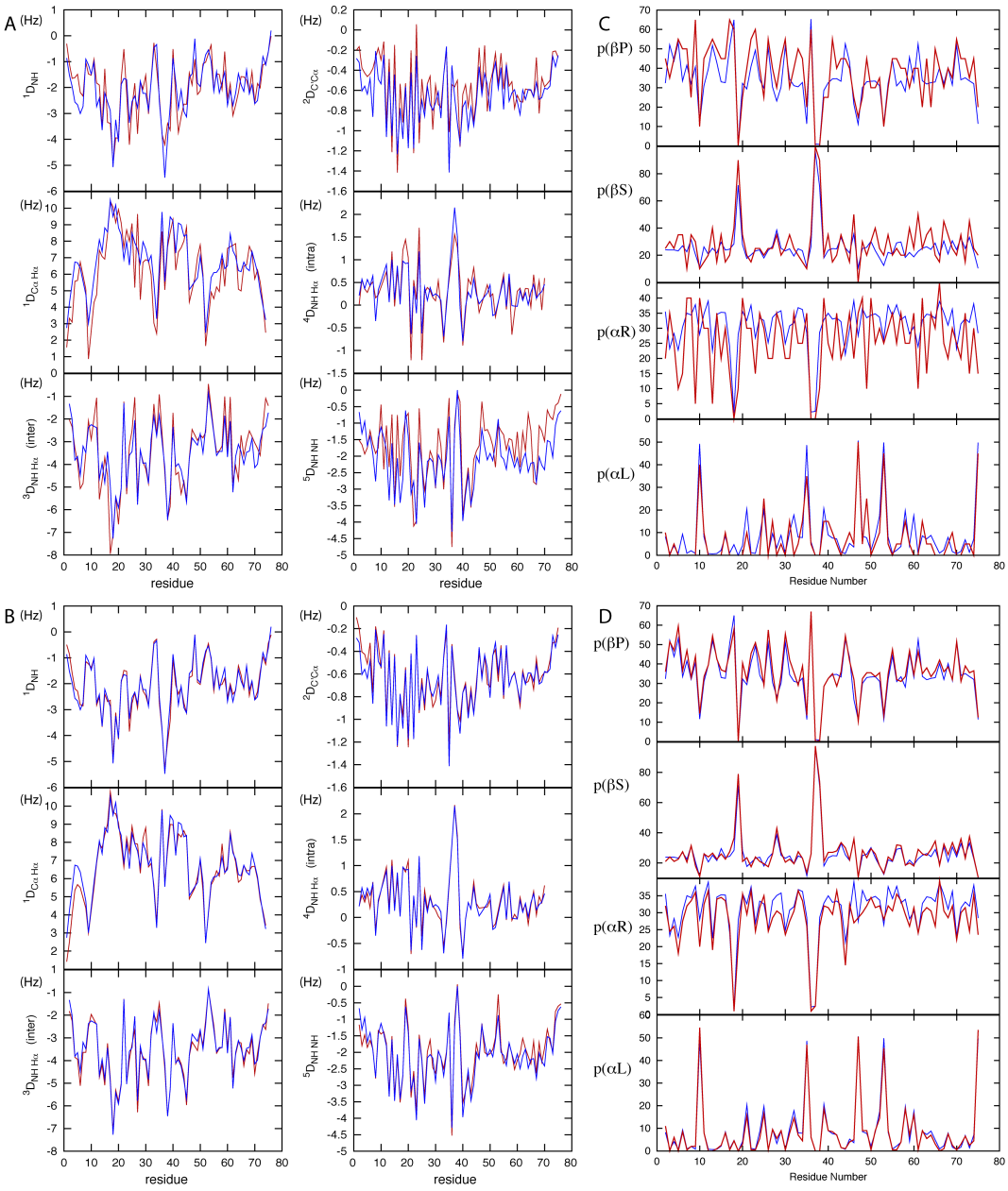

Figure 71 – Site specific RDCs (A-B) and conformational sampling (C-D) reproduction of simulated data using the ASTEROIDS approach: (A) and (C) 20-fold (B) and (D) 200-fold ensembles. (A) and (B) ${}^{1}D_{NH}$, ${}^{1}D_{C^{\alpha}H^{\alpha}}$, ${}^{1}D_{C'C^{\alpha}}$, ${}^{3}D_{H^{N}H^{\alpha}}$, ${}^{4}D_{H_i^N H_{i-1}^{\alpha}}$ and ${}^{5}D_{H_i^N H_{i+1}^N}$ couplings : simulated target (blue), ASTEROIDS selected ensemble (red). (B) population of the different quadrants: conformational sampling obtained with ASTEROIDS selected ensemble (red) compared to the targeted distribution (blue).

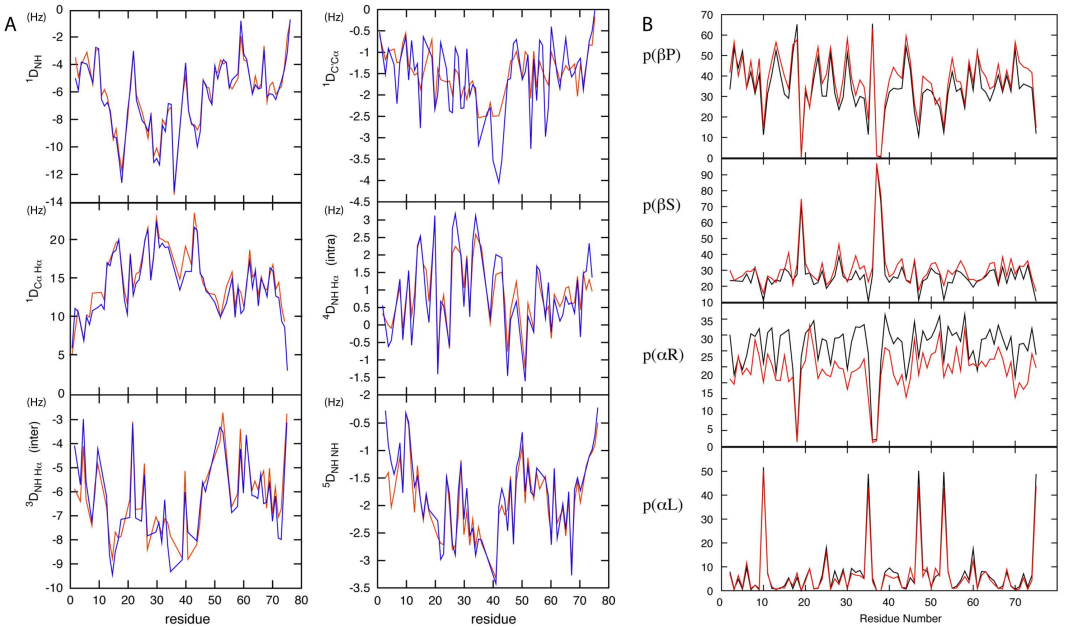

Figure 72 – (A) $^{1}D_{NH}$, $^{1}D_{C^{\alpha}H^{\alpha}}$, $^{1}D_{C'C^{\alpha}}$, $^{3}D_{H^{N}H^{\alpha}}$, $^{4}D_{H_i^N H_{i-1}^{\alpha}}$ and $^{5}D_{H_i^N H_{i+1}^N}$ RDCs reproduction of urea-denatured Ubiquitin data using the ASTEROIDS approach with differential scaling of the covalently bound and inter-protons RDCs (red) compared to experimental data (blue). (B) obtained conformational sampling of urea-denatured Ubiquitin (red) compared to the random-coil distribution (black).

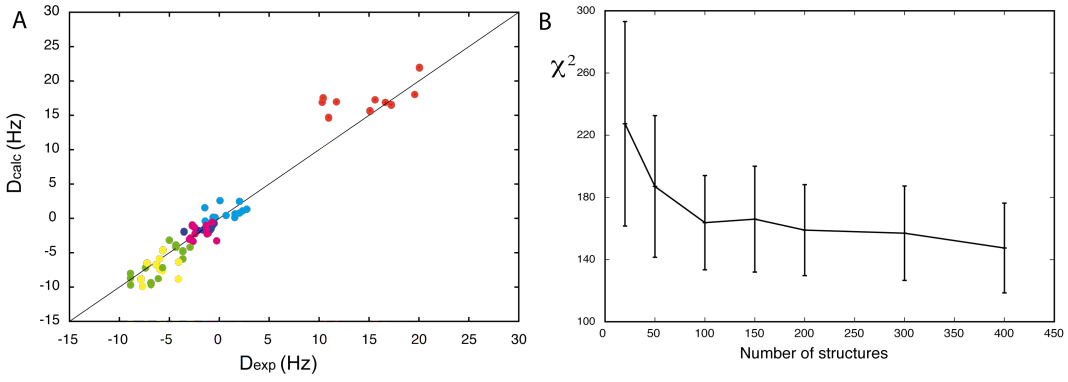

Figure 73 – (A) Cross-validation of the urea-denatured Ubiquitin analysis using the ASTEROIDS approach. Passive data reproduction compared to experimental RDCs. $^{1}D_{NH}$ (green), $^{1}D_{C^{\alpha}H^{\alpha}}$ (red), $^{1}D_{C'C^{\alpha}}$ (dark blue), $^{3}D_{H^{N}H^{\alpha}}$ (cyan), $^{4}D_{H_i^N H_{i-1}^{\alpha}}$ (yellow) and $^{5}D_{H_i^N H_{i+1}^N}$ (magenta). (B) Passive data reproduction averaged over 10 cross-validation calculations as a function of the ensemble size.

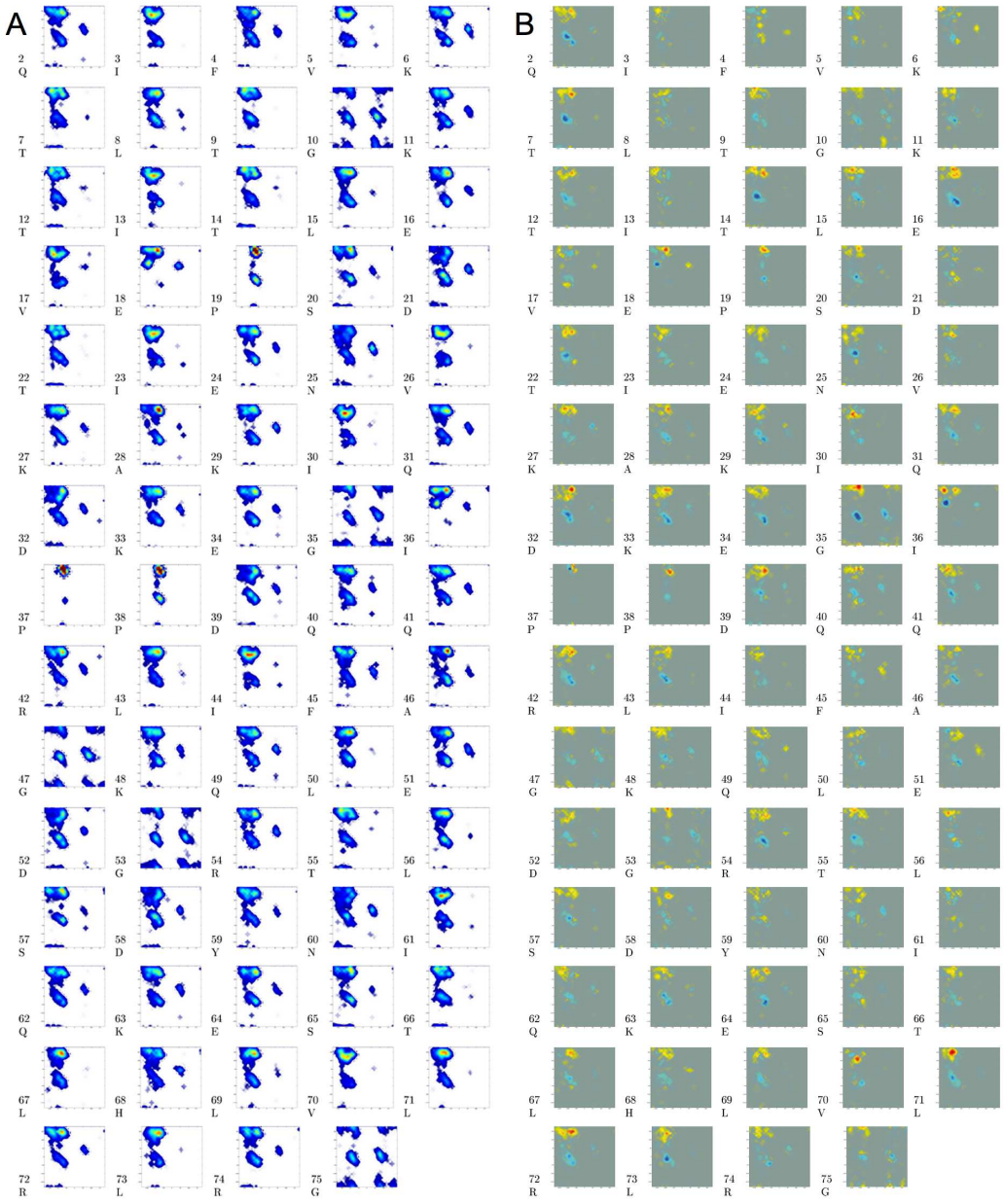

Figure 74 – Amino-acid specific Ramachandran distributions for urea-denatured Ubiquitin compared to a random-coil sampling. (A) Conformational sampling determined from the ASTEROIDS analysis (10 calculations were combined to reach 2000 conformers to increase resolution). Populations increase from dark blue, via cyan, green, yellow to red. (B) Difference between the conformational sampling distributions obtained for urea-denatured Ubiquitin and the random-coil distribution. Color scale: blue to green corresponds to negative values (population lower for urea-denatured Ubiquitin), and yellow to red corresponds to positive values (population higher for urea-denatured Ubiquitin). Green corresponds to equal populations in both cases.

adopts more extended conformations compared to the standard random-coil sampling. This can be due to their capability to interact with urea through hydrogen-bonds. The fact that at low pH, urea orients preferentially towards the protein [270] supports the idea that urea interacts preferentially with hydrogen-bond donor residues. By contrast hydrophobic amino-acids seem to exhibit only minor differences between the urea denatured state and the random-coil. A small angle scattering study [271] proposed that around 20 urea molecules per protein are directly involved in the denaturation, which is qualitatively in agreement with this study as approximatively one third of the amino-acids exhibit important deviations from the random-coil distribution.

10.4 CONCLUSION FOR UREA-DENATURED UBIQUITIN ANALYSIS

This study presented a novel approach that can be used to select conformational ensembles of disordered states on the basis of extensive experimental residual dipolar couplings. We find that the particular averaging properties of dipolar couplings, and their sensitivity to both local conformational sampling and the chain-like nature of the disordered state, imposes specific requirements on the development of ensemble descriptions from these particular data sets. One consequence is that many tens of thousands of conformers are required to correctly average RDCs calculated over the entire chain dimensions of a protein of say 100 amino-acids. Another, related consequence, is that, due to the immense number of degrees of conformational freedom compared to experimental constraints, data can be well reproduced by ensembles whose conformational properties do not accurately represent the true behavior of the protein, if applied in the inappropriate averaging regime.

We provide solutions to these problems, using a local alignment window, describing the reorientational properties of a 15 amino-acid segment around the amino-acid of interest, thereby providing amino-acid specific conformational properties of each amino-acid. We note however that the chain nature of the protein is thereby ignored, and this has to be corrected by introducing a contribution due to the polymer chain.

Extensive simulations demonstrate that between 100 and 200 structures are then necessary to correctly describe the conformational behavior when using 15 amino-acid alignment segments. This provides accurate amino-acid sampling along the chain. The use of fewer structures can again reproduce the data, but provides an incorrect conformational description.

Having developed the selection algorithm ASTEROIDS, that can in principle be applied to selection on the basis of any experimental parameter that can be calculated for each conformer and then averaged over the ensemble, the most obvious application was to apply the approach to the interpretation of chemical shifts. In the next Section we describe how we have combined recent developments in the prediction of chemical shifts from individual structures to describe the conformational properties of partially folded chains on the basis of chemical shifts alone.

10.5 INTRODUCTION FOR N_{TAIL} ANALYSIS USING CHEMICAL SHIFTS

Chemical shifts measured in IDPs report on the population weighted average over an entire ensemble of interchanging conformers, exchanging on timescales faster than the millisecond range. These readily measured parameters are nevertheless highly sensitive probes of local protein conformation, as has been demonstrated by the recent determination of three-dimensional structures of entire globular proteins using chemical shifts as sole experimental constraints. The dependence of ^{13}C backbone chemical shifts on backbone (ϕ/ψ) dihedral angles, have been routinely used to identify secondary structure, and to estimate the level of secondary structural propensity within folded and unfolded proteins.

In this study we combine ensemble descriptions of unfolded proteins, with state-of-the-art chemical shift prediction algorithms that have underpinned the successful determination of folded proteins from chemical shifts. This combination is used to explore the possibility of using chemical shifts alone to map local backbone conformational sampling of intrinsically disordered and partially folded proteins.

10.6 MATERIALS AND METHODS FOR N_{TAIL} ANALYSIS

10.6.1 *Experimental Data*

N_{TAIL} Chemical Shifts were measured by Jensen et al. [261]. Potential chemical shift reference offsets were corrected using SSP [237]. Secondary Chemical Shifts (SCSs) were calculated as the difference between the experimental CSs and the random-coil values from RefDB [272], except for residues preceding prolines where random-coil values were taken from Wishart et al. [233].

10.6.2 FLEXIBLE-MECCANO *Ensemble Generation, CSs Calculations and* ASTEROIDS *Selection*

Similar protocol than for the urea-denatured Ubiquitin study was used (see Section 10.2.2). The starting database consisted of 10 000 N_{TAIL} structures obtained using the standard random-coil FLEXIBLE-MECCANO sampling. For each conformer the side-chains were added using SSCOMP [273] and the CSs were calculated for each conformer using a modified version of SPARTA [274] that did not include the side-chain χ_1 torsion angle in the scoring function as no information is available about the exact conformations of the side chains.

For ASTEROIDS selection, the weights of the CSs were 2.0 for ^{15}N and 1.0 for $^{13}C'$, $^{13}C^{\alpha}$ and $^{13}C^{\beta}$. Two kinds of analyses were successively carried out: one where each amino-acid is treated independently and one with selection of full structures on the basis of the experimental CSs.

For the treatment of individual amino-acids, the ^{15}N CS was not used. For each residue, 200 structures are selected and the analysis is repeated five times in order to reach 1000 (ϕ, ψ)-combinations. 500 generations are enough to obtain convergence. The selection is used as a modified (ϕ, ψ)-database for FLEXIBLE-MECCANO in order to generate a pool of 7500 structures. Note that at this point the (ϕ, ψ)-database is different for each amino-acid in the protein. The pool of 7500 structures is complemented with 2500 structures obtained using the standard random-coil distribution. This procedure ((ϕ, ψ)-combination selection and new FLEXIBLE-MECCANO structure generation) is repeated until no further improvement in experimental data reproduction is achieved.

A last step involves selecting 200 conformers from the last obtained pool of structures that agree with all experimentally measured CSs of the protein. In this step the ^{15}N CSs are included and 3,000 generations are necessary to ensure convergence.

10.7 RESULTS AND DISCUSSION FOR N_{TAIL} CHEMICAL SHIFTS ANALYSIS

10.7.1 *Ability of the* ASTEROIDS *Protocol to Define Conformational Sampling*

The FLEXIBLE-MECCANO-ASTEROIDS protocol was tested on simulated CS data in a similar way to the urea-denatured Ubiquitin approach. Three sets of simulated CS data corresponding to FLEXIBLE-MECCANO ensembles created using a standard random-coil database, a more extended database

and a more helical database were submitted to ASTEROIDS analysis. The targeted data corresponding to the extended and helical ensembles could not be closely reproduced by selecting full structures using ASTEROIDS from a large pool of conformers created using the standard random-coil database. An explanation for this resides in the inadequacy of the standard random-coil database to provide a sufficient number of structures with high proportions of e.g. helical propensity.

A potential solution to this problem is to heavily over-sample i.e. to increase the size of the FLEXIBLE-MECCANO pool however this process will be very time consuming. Thus, a modification of the FLEXIBLE-MECCANO (ϕ, ψ)-database was chosen. A new (ϕ, ψ)-database adapted to the protein under investigation is obtained by pre-selecting for each amino-acid in the protein a smaller (ϕ, ψ)-database through the iterative process presented in Section 10.6.2. For the individual treatment ^{15}N CSs were not used as they mainly depend on the conformation of the neighboring residues (mainly the ϕ-angle of the previous amino-acid) [274]. After typically three or four iterations no more improvement in the data reproduction was observed.

Using this site-specific (ϕ, ψ)-database, 10 000 FLEXIBLE-MECCANO structures were generated and ASTEROIDS selection of 200 structures was achieved. Good data reproduction was obtained and the conformational sampling used to simulate the data was reproduced within 5% accuracy.

10.7.2 *Conformational Sampling of N_{TAIL} from Chemical Shifts*

The protocol was applied to N_{TAIL} experimental data. This model system was chosen, as it was already demonstrated using a RDC based approach that the central part of the protein adopts fluctuating helical elements. Four iterations were necessary to obtain a converged residue specific database. The data reproduction obtained with the final 200 extracted structures is shown in Figure 75. The overall agreement is very good, validating the ability of the protocol to reproduce experimental data.

Nevertheless the physical relevance of the obtained ensemble has to be tested. Therefore, the conformational sampling obtained directly from the CSs was used to back-calculate independent experimental data on the same system, namely the $^1D_{NH}$ RDCs measured in a PEG/hexanol mixture [261]. The RDCs were calculated using 50 000 FLEXIBLE-MECCANO conformers generated using the same (ϕ, ψ)-database as the 200 selected structures on the basis of the CSs. As the alignment is of sterical origin, RDCs were calculated using PALES and a scaling factor was applied to reproduce the absolute level of alignment. Another validation test consisted

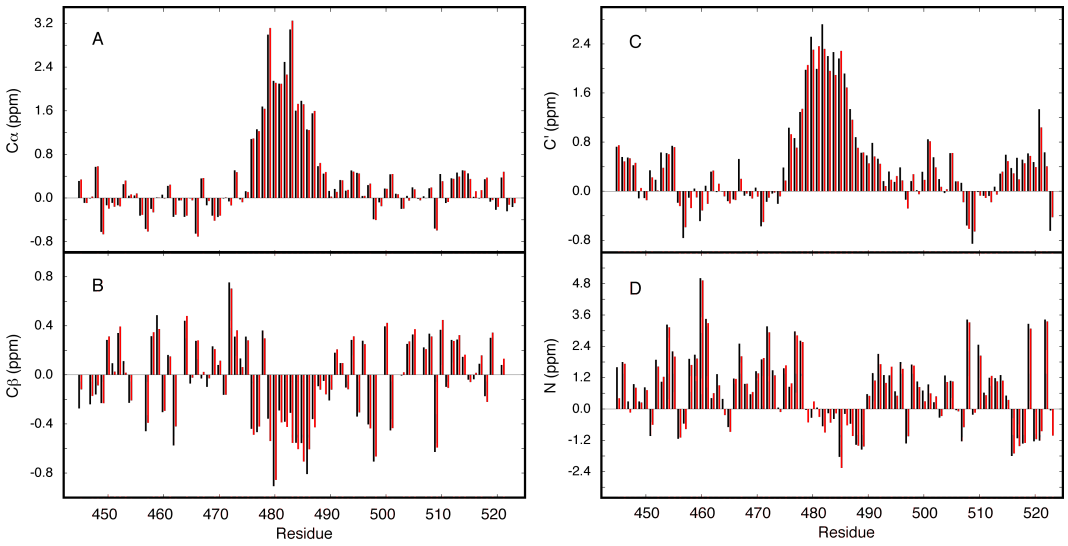

Figure 75 – Reproduction of N_{TAIL} experimental secondary chemical shifts (black) with the ASTEROIDS selected ensemble (red). (A) $^{13}C^{\alpha}$, (B) $^{13}C^{\beta}$, (C) $^{13}C'$ and (D) ^{15}N.

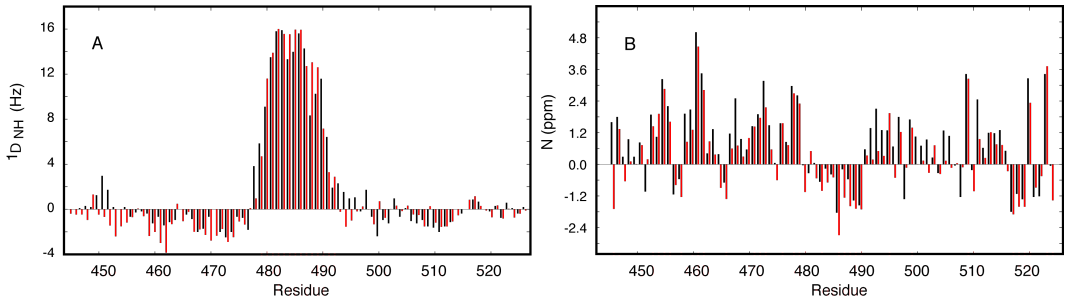

Figure 76 – Reproduction of independent data (black) by the ASTEROIDS obtained ensemble (red): (A) $^{1}D_{NH}$ couplings (B) ^{15}N secondary chemical shifts.

of removing all ^{15}N CSs and repeating the complete ASTEROIDS analysis and subsequently looking at the capability of the selected conformational sampling to reproduce the ^{15}N CSs. Both results are shown in Figure 76.

The agreement is very good for the $^{1}D_{NH}$ couplings. Even if some details are not matched, the major features are accurately reproduced. This result is a good validation of the approach, as RDCs are very sensitive probes of conformational sampling and they represent completely independent data. The data reproduction of the ^{15}N CSs is a bit worse than the $^{1}D_{NH}$ couplings, but in this case a quarter of the data was removed from the analysis and therefore the robustness of the approach is compromised. Nevertheless the back-calculated ^{15}N CSs from this indirect analysis clearly better reproduce experimental data than back-calculated shifts from the standard random-coil distribution (root mean square deviation of 0.77 ppm

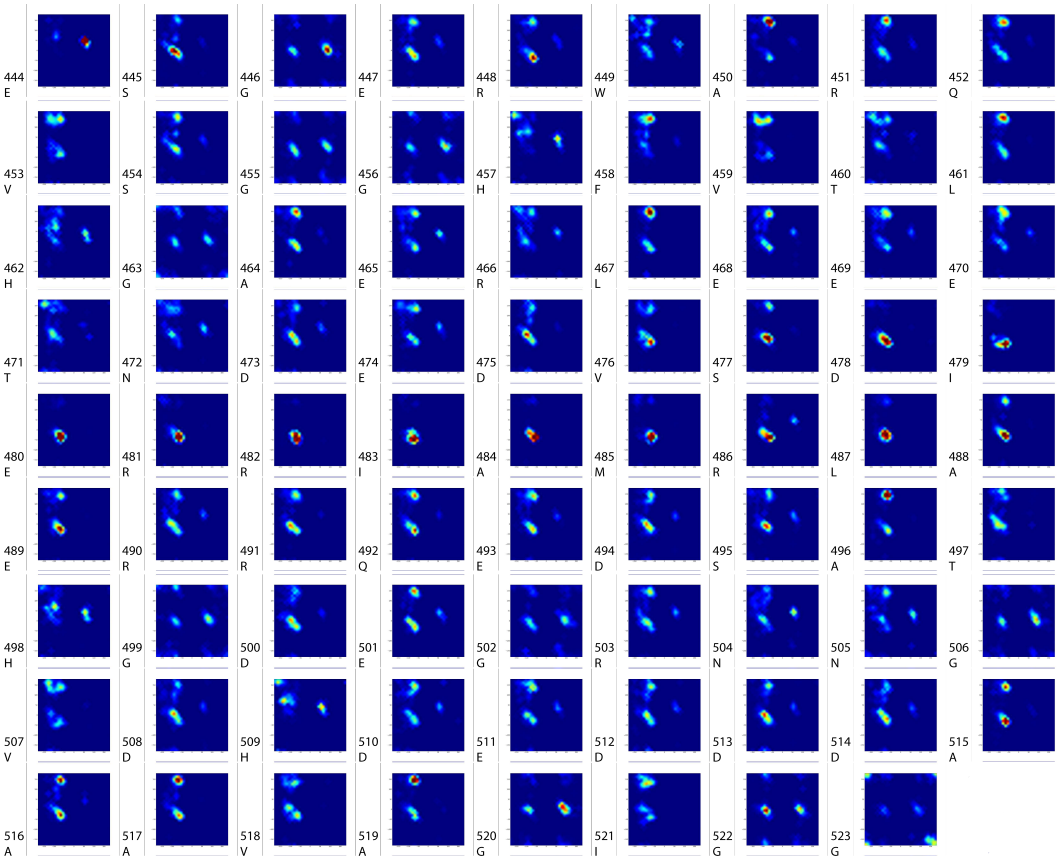

Figure 77 – Amino-acid specific Ramachandran distributions for N_{TAIL}. Populations increase from dark blue, via cyan, green, yellow to red. Conformational sampling determined from the Asteroids analysis (10 calculations were combined to reach 2,000 conformers to increase resolution).

compared to 1.15 ppm for the random-coil model). Both results support the validity of the protocol.

Propensity to form α-helix in the center of the protein is clearly visible (see Figure 77) which is in good agreement with the results obtained from an extensive analysis of the N_{TAIL} RDCs [261]. Outside the helical region, a tendency to populate more the β_P region instead of the β_S region was found. This is in agreement with complementary approaches that suggest that poly-Proline conformations (β_P) are dominant in IDPs [275, 276].

10.8 CONCLUSION

In this study the Asteroids methodology was developed in order to obtain insight into local conformational sampling in IDPs on the basis of experimental backbone CSs. The Asteroids approach allowed the extraction of

a representative sub-ensemble that agrees with experimental data from a large pool of structures without using any hypotheses *a priori*.

For the N_{TAIL} study it was possible to find the propensity of α-helix formation. The residual order present in this system has been shown to be particularly important for initiating interaction with its biological partner [261, 277]. The application of the ASTEROIDS approach in the future to other unfolded systems would allow detection of such biologically important features.

In the case of urea-denatured Ubiquitin, a detailed site-specific characterization of the conformational sampling was obtained. The sampling clearly deviates from that of the standard random-coil state revealing the effect of urea on the conformational properties of the unfolded chain.

The information content of the CSs and RDCs is slightly different, where the latter is sensitive to both local and long-range order. CSs provide an experimentally easier way to characterize the site specific conformational sampling, but the obtained information depends mainly on the amino-acid of interest and the adjacent neighbors. On the contrary, RDCs depend on both local information (using the LAW tensor determination) and global information (using the baseline parameterization). In the analyses presented here it was assumed that no long-range order was present as the applied RDC baseline was obtained from the poly-Valine created using the standard random-coil database. The absence of significant long-range order in urea-denatured Ubiquitin is in agreement with previous studies [265]. However, a description is necessary of the effect of long-range order on measurable NMR parameters such as RDCs.

11

CHARACTERIZATION OF LONG-RANGE ORDER IN UNFOLDED SYSTEMS

ABSTRACT

Even in unfolded systems long-range order, albeit transient, can be present in solution. The potential manifestation and effects of this order for different experimental parameters are investigated. The novel approach presented in the previous chapter for the description of conformational ensembles is tested on simulated data to demonstrate the ability of the protocol to identify ill-defined contacts under the application of experimental PREs of α-Synuclein. This allows the detection of the spatial proximity between N- and C-terminal domains of the protein. The influence of long-range order on RDCs is investigated and parameterized. Using the parameterization obtained from contacts detected using PREs, the reproduction of α-Synuclein RDCs is clearly improved compared to a description where no long-range order is supposed.

11.1 INTRODUCTION

The study of IDPs has shown that they do not always follow the perfect random-coil description. Most of the studies focus on a local interpretation of the data e.g. in terms of the propensities for forming secondary structures [230, 261, 266]. The characterization of residual structure in unfolded proteins is particularly important as regions with secondary structures are often implicated in interactions with partner proteins.

As described in the previous chapter, the description of RDCs can be done by separating local and long-range order. The chain like behavior of the IDPs is modeled in terms of an appropriate hyperbolic cosine function and the local properties of the RDCs are described using Local Alignment Windows. This mainly corresponds to treating explicitly the influence of the near neighbors and use an over-all description for the chain-like nature of the protein.

In the previous study, the applied baseline was inspired by a completely unfolded system. Intramolecular interactions can however of course exist in IDPs between different parts of the protein, for example between complementarily charged fragments or between hydrophobic regions. It is probable that the effect of such interactions on RDCs has to be explicitly taken into account in order to avoid interpreting such long-range order in terms of local conformational sampling.

PREs are very powerful probes of distance distributions and long-range order in IDPs. The aim of this chapter is to derive an appropriate description of long-range order in the unfolded state from PRE, to characterize its influence on RDCs and thereby to combine the two highly complementary sources of conformational information in a single analysis. The approach will be developed using α-Synuclein (see Annexe B) for which PREs have already been measured for four Cysteine mutants with attached MTSL spin-labels [252].

11.2 THEORY AND METHODS

11.2.1 *Dynamic Averaging of PREs in* FLEXIBLE-MECCANO *Ensemble Description*

The flexibility of the considered system, an IDP with an MTSL spin-label attached to a cysteine residue, is described here using the combination of two different dynamical models. The first model takes into account the exchange between different conformers occurring in highly dynamical systems such as IDPs, whereas the second model takes into account the mobility of the MTSL spin label on each conformer. It is assumed that the FLEXIBLE-MECCANO ensemble description represents the dynamics of the protein backbone. The two motions are assumed to occur on different timescales and therefore to be statistically independent [46]. Even if the precise timescales are not known, the interconversion between the FLEXIBLE-MECCANO conformers has to occur at timescales faster than the milli-second and can be expected to be on the ns-μs timescales as geometrical reorganization of the backbone can be significant between different conformers, whereas the dynamics of the MTSL spin label is expected to occur at the picosecond timescale (ten to hundreds of picoseconds) as it includes a small number of side chain reorientations. The MTSL motion is modeled by sampling a rotameric distribution developed by Sezer et al. [278] and retaining conformations that do not have steric clashes with the protein backbone. This flexibility is illustrated in Figure 78. For each FLEXIBLE-MECCANO conformer k, the contribution to the transverse relaxation rate at residue i due to the presence

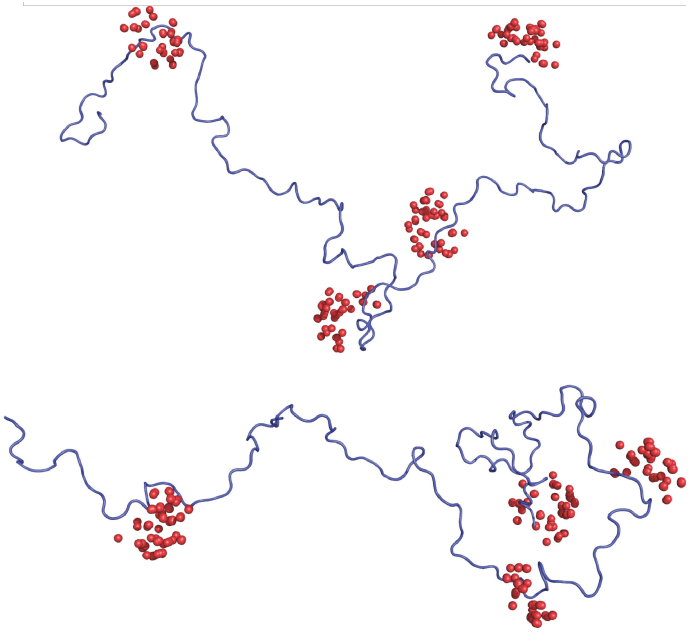

Figure 78 – Representation of the possible nitroxide spin-label positions (red dots) for two FLEXIBLE-MECCANO conformers (blue ribbon) for the four paramagnetic probes used in the α-Synuclein study (amino-acids 18, 76, 90, and 140).

of the unpaired electron anchored at residue m, $\Gamma_{2,k,i,m}$, can be expressed as [279, 280]:

$$\Gamma_{2,k,i,m} = \frac{2}{5}\left(\frac{\mu_0 \gamma_H g_e \mu_B}{4\pi}\right)^2 S_e(S_e+1)\left[4J(0)+3J(\omega_H)\right] \tag{11.1}$$

where g_e is the Landé electron g-factor, μ_B the Bohr magneton and S_e the electron spin.

Using a previously developed expression for homonuclear cross relaxation, it has been shown [250, 281, 282] that the spectral density function J corresponding to the electron-nucleus interaction can be expressed in a model-free way, where an order parameter is invoked that depends on both the orientation and length of the electron-nucleus vector:

$$J(\omega) = \left\langle r_{H-e}^{-6} \right\rangle \left[\frac{S_{H-e}^2 \tau_r}{1+\omega^2\tau_r^2} + \frac{(1-S_{H-e}^2)\tau_e}{1+\omega^2\tau_e^2}\right] \tag{11.2}$$

where S_{H-e}^2 is the order parameter of the electron-nucleus vector, r_{H-e} is the instantaneous electron-nucleus distance and the correlation times are given by [281, 282]:

$$\frac{1}{\tau_r} = \frac{1}{\tau_C} + \frac{1}{\tau_s} \qquad \text{and} \qquad \frac{1}{\tau_e} = \frac{1}{\tau_C} + \frac{1}{\tau_s} + \frac{1}{\tau_I} \tag{11.3}$$

where τ_C is the correlation time of the molecule, τ_s is the electron spin relaxation time and τ_I is the correlation time of the electron-nucleus motion.

The order parameter S^2_{H-e} can to a good approximation be decomposed in an angular and a radial contribution [281]:

$$S^2_{H-e} = \frac{4\pi}{5} \left\langle r^{-6}_{H-e} \right\rangle^{-1} \sum_{m=-2}^{2} \left| \left\langle \frac{Y_{2,m}(\theta,\phi)}{r^3_{H-e}} \right\rangle \right|^2 \simeq S^2_{ang} S^2_{rad} \tag{11.4}$$

where

$$S^2_{rad} = \left\langle r^{-3}_{H-e} \right\rangle^2 \left\langle r^{-6}_{H-e} \right\rangle^{-1} \qquad \text{and} \qquad S^2_{ang} = \frac{4\pi}{5} \sum_{m=-2}^{2} \left| \langle Y_{2,m}(\theta,\phi) \rangle \right|^2 \tag{11.5}$$

in which (θ, ϕ) represent the orientation of the electron-nucleus vector in the molecular frame.

These equations allow an estimation of the contribution from the unpaired electron spin to the transverse relaxation rate for a given conformer. For an ensemble of N conformers, the effective contribution is:

$$\Gamma_{2,i,m} = \frac{1}{N} \sum_{k=1}^{N} \Gamma_{2,k,i,m} \tag{11.6}$$

11.2.2 *PRE and RDC Calculations from* Flexible-Meccano *Conformers*

For a Flexible-Meccano ensemble of conformers, the PREs are estimated using the formalism presented in Section 11.2.1. For this study, the values of τ_r was fixed to 5 ns, as proposed in the literature [140, 263]. The effective correlation time for the MTSL reorientation τ_e was set to 500 ps. This value is in broad agreement with Electron Spin Relaxation studies [278], and changing it by a factor of two in either direction does not lead to noticeable differences in the obtained results. For the MTSL dynamics, 600 conformers of the side chain were randomly generated using the MTSL rotameric library proposed previously, however, only conformers without sterical clashes with the backbone were retained.

Calculation of RDCs can be done using a global alignment tensor or combination of using an appropriate baseline and the LAW description (see Section 10.2.2). Briefly, a global alignment tensor corresponds to estimating the tensor using Pales applied to the full length protein and to average the results over all conformers. The LAW description uses a 15 amino-acid window to estimate a local tensor for each residue in the protein. The influence of the rest of the chain on the RDCs is taken into account by using a so-called baseline. The baseline is a function of the position of the amino-acid in the peptide chain, but also depends on the existence of contacts between distant parts of the chain (see Section 11.2.6).

11.2.3 ASTEROIDS *Ensemble Selection*

The ASTEROIDS ensemble selection is carried out as described in Section 10.2.4. The number of individuals per generation P is fixed to 100 and 2,000 generations were used. The fitting function can be expressed as follows:

$$\chi^2 = \sum_{i,m} \left(\left[\frac{I_{para}}{I_{dia}} \right]_{i,m}^{calc} - \left[\frac{I_{para}}{I_{dia}} \right]_{i,m}^{exp} \right)^2 \tag{11.7}$$

where I_{para} and I_{dia} are the intensity of a resonance in the paramagnetic (oxidized nitroxide) and diamagnetic (reduced nitroxide) form of the MTSL. This expression corresponds to using identical weights for all measurements. The intensity ratio, at residue i with MTSL attached at residue m is estimated using [282]:

$$\left[\frac{I_{para}}{I_{dia}} \right]_{i,m}^{calc} = \frac{R_{2,i}\, e^{-\Gamma_{2,i,m}\tau_{mix}}}{R_{2,i} + \Gamma_{2,i,m}} \tag{11.8}$$

where $R_{2,i}$ is the intrinsic relaxation rate of the amide proton of the amino-acid i, τ_{mix} is the mixing time of 10 ms where relaxation occurs (corresponding to the INEPT — Insensitive Nuclei Enhanced by Polarization Transfer — period of an HSQC) and $\Gamma_{2,i,m}$ is the contribution to relaxation due to the presence of the paramagnetic center, estimated as described in Section 11.2.1.

11.2.4 *Contact Definition*

Contacts are defined in a broad way. For generating ensembles of structures, a contact between two regions is considered to occur if at least one C^β of the first contiguous strand (e.g. residues 11-20) is located less than 15 Å from any C^β in the second contiguous strand (e.g. residues 51-60). The two regions has to be separated by at least 20 amino-acids in order to avoid over-interpreting sequence proximity in terms of spatial vicinity.

11.2.5 *Contact Matrices*

Average distances between two residues i and j were represented using a Δ_{ij} metric defined as:

$$\Delta_{ij} = \log \frac{\left\langle d_{ij} \right\rangle}{\left\langle d_{ij}^0 \right\rangle} \tag{11.9}$$

where $\left\langle d_{ij} \right\rangle$ is the average distance in the considered ensemble and $\left\langle d_{ij}^{0} \right\rangle$ is the average distance in the reference ensemble. The reference ensemble consists of 10,000 FLEXIBLE-MECCANO conformers from which no specific selection was made i.e. the ensemble does not contain any favored long-range interactions. Distances d_{ij} are calculated as the distance between the C^{α} of residue i and the C^{α} of residue j.

This metric was used as a probe for the presence of contacts within a given ensemble. It is worth noting that this representation enhances contacts that are further apart in the chain and thus the observed contacts tend to be "smeared" away from the diagonal in contact maps.

The determination of the dominant contact is done by first locating the maximal discrepancy between the selected ensemble and the reference ensemble, i.e. Δ_{ij}^{max}. The matrix is then divided in 5×5 amino-acid regions. The region with the highest number of pairs of residues fulfilling:

$$0.9\,\Delta_{ij}^{max} < \Delta_{ij} < \Delta_{ij}^{max} \tag{11.10}$$

is identified as the contacting region and the center coordinates of this region are retained for use in the baseline expression discussed below.

11.2.6 *Baseline Parameterization*

The generic baseline taking into account the influence of long-range contacts on RDCs is an extension of the previously proposed hyperbolic cosine function (see Chapter 10) used in the case where no specific contacts occur. The effect of the contact is modeled by a Gaussian centered between the two contacting segments and two smaller Gaussians are added near the interacting regions in order to correct the local curvature. The parameterization of the Gaussian depends only on the length of the peptide chain L and on the position of the two interacting regions n_1 and n_2:

$$\begin{aligned} B(i, L, n_1, n_2) = &\left[2\,b\,\cosh\left(a(m - m_0) \right) - c \right] \\ &\times \left[1 - G\,e^{-\frac{(m-n_0)^2}{2\sigma^2}} + H \right. \\ &\left. \times \left[(D+S)\,e^{-\frac{(m-n_1+S/2))^2}{2\delta^2}} + (D-S)\,e^{-\frac{(m-n_2-S/2))^2}{2\delta^2}} \right] \right] \end{aligned} \tag{11.11}$$

where m_0, a, b and c are functions of the length L of the chain:

$$\begin{aligned} m_0 &= \frac{L+1}{2} & a &= 0.33 - 0.22\left[1 - e^{-0.015\,L}\right] \\ b &= 1.16\,10^5\,L^{-4} & c &= 9.80 - 6.14\left[1 - e^{-0.021\,L}\right] \end{aligned} \tag{11.12}$$

where n_0, D, σ, S, G and H are functions of the two positions of the interacting regions n_1 and n_2:

$$\begin{aligned} n_0 &= \frac{n_1+n_2}{2} & D &= \left|n_1 - n_2\right| & \sigma &= 0.109\,D + 4.6\,10^{-3}\,D^2 \\ S &= n_0 - m_0 & H &= 3.87\,10^{-5}\,D & G &= 1 - 6.66\ 10^{-3}\,D \end{aligned} \tag{11.13}$$

and $\delta = 9.0$. As a contact is defined between two contiguous strands, the n_1 and n_2 positions are fixed at the center of the two interacting strands.

This expression depend on the length of the protein L and the position of the two contacts n_1 and n_2. The i-th coupling of between spins I and S can be calculate as:

$$D_{IS}(i) = |B(i, L, n_1, n_2)| D_{IS}^{LAW} \tag{11.14}$$

11.2.7 *Radius of Gyration Calculations*

The radius of gyration is calculated as presented in Section 10.2.6.

11.2.8 *Experimental Data*

Experimental data were measured by Zweckstetter and co-workers. PREs were measured on four different mutants where Alanines where substituted by Cysteines: mutants A18C, A76C, A90C and A140C. RDCs were measured in PEG/hexanol mixtures. Experimental details can be found in references [252, 283].

11.2.9 *Simulated Data*

PREs were simulated for a 100 amino-acid protein of arbitrary sequence by averaging the PREs over 10,000 Flexible-Meccano conformers. Simulated data were obtained for four different positions of the MTSL spin label (at residues 20, 40, 60 and 80) where contacts between the regions 11-20 and 61-70 or between 41-50 and 81-90 were imposed.

PREs were also simulated for a 200 amino-acid protein of arbitrary sequence by averaging the PREs over 10,000 FLEXIBLE-MECCANO conformers. Simulated data were obtained for eight different positions of the MTSL spin label (at residues 22, 44, 66, 88, 110, 132, 154 and 176) where contacts between the regions 11-20 and 61-70 or between 141-150 and 181-190 were imposed.

11.3 RESULTS AND DISCUSSION

11.3.1 *Testing* ASTEROIDS *Approach on PRE Simulated Data*

The ASTEROIDS approach was tested using the simulated data for the 100 amino-acid model protein (see Section 11.2.9). Figure 79 shows the targeted simulated data, the reproduction of those data with an ensemble of 80 structures selected using ASTEROIDS and the profile obtained from the database consisting of 10,000 conformers with no specific contacts from which the ASTEROIDS ensemble was selected.

Even in absence of specific contacts, the impact of the spin-label presence is felt in a large vicinity of the residue where the MTSL is grafted. The data reproduction obtained with the ASTEROIDS ensemble is very good, even if the targets are quite demanding as 20% of the protein is involved in weak contacts. These simulations nevertheless represent a reasonable reproduction of the situation that one may encounter when studying intrinsically disordered or partially folded proteins, with long-range interactions occurring between strands carrying complementary electrostatic charge or containing hydrophobic side chains.

The biophysical relevance of the obtained ASTEROIDS ensembles is investigated by comparing the obtained ensembles with the target ensembles through the distribution of distances (Figure 80) and the distribution of radii of gyration (Figure 81).

The contact maps of the selected ASTEROIDS ensemble and the target ensemble are very similar although the contact in the case of the ASTEROIDS ensemble seems a bit less well-defined compared to the target ensemble. The exact values of the distances were not reproduced (the distances were underestimated), but this is not considered as a serious drawback in view of the ill-defined nature of the contact.

Concerning the radii of gyration R_g, the distribution is qualitatively well reproduced. For the selected ASTEROIDS ensemble (80 structures) the average R_g is 21.3 Å compared to 22.6 Å for the target ensemble. Selection of an ensemble with 160 conformers lead to an increase of R_g to 21.7 Å. The

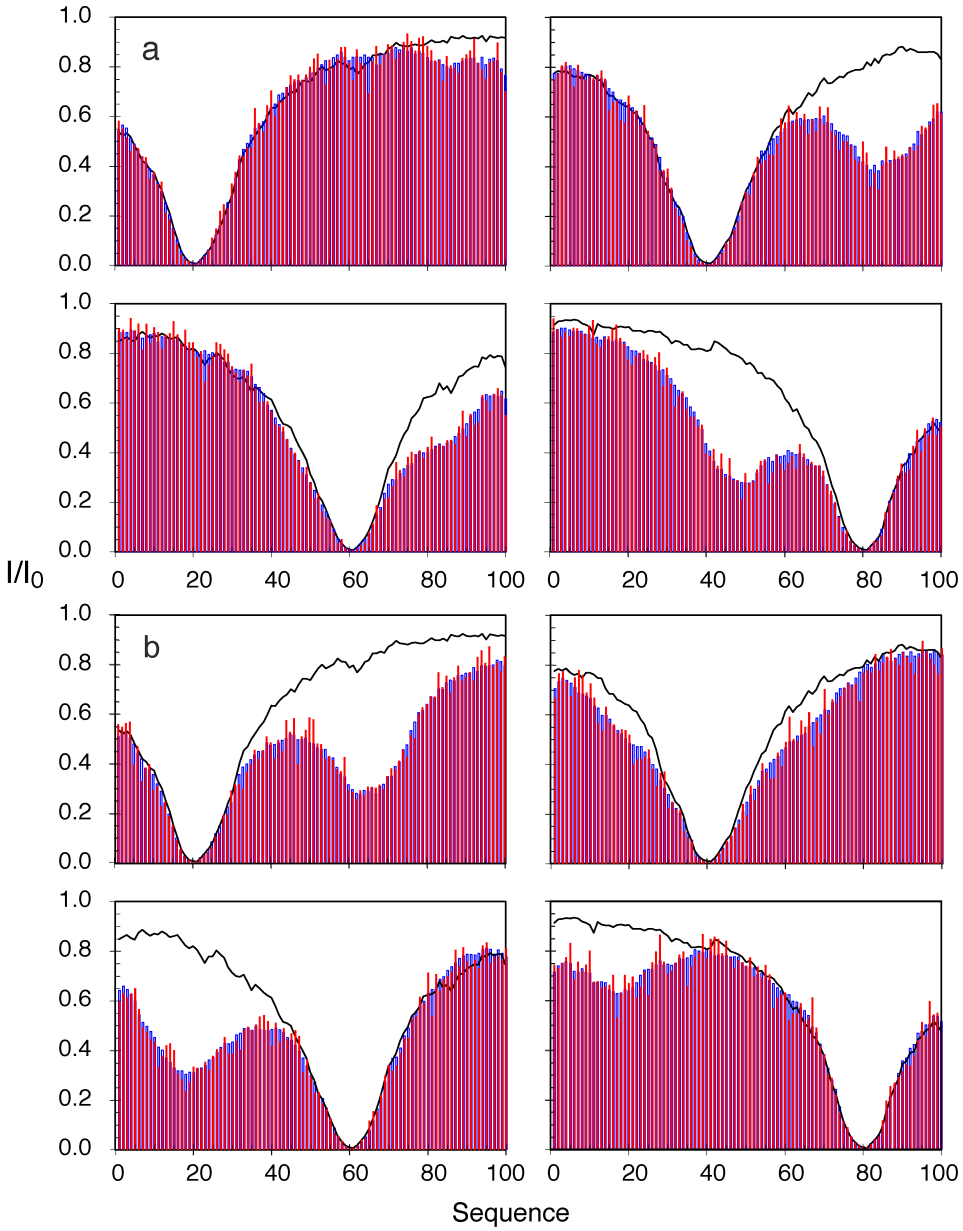

Figure 79 – Reproduction of simulated PREs for ensembles containing specific contacts using ASTEROIDS. Target ensembles with a contact between 41-50 and 81-90 (a) or between 11-20 and 61-70 (b). Blue: data averaged over the target ensemble. Red: data averaged over an ASTEROIDS selected ensemble. Lines show the PREs calculated from an ensemble with no specific contacts. Each box shows the PRE data for a spin-label at position 20 (top left), 40 (top right), 60 (bottom left), and 80 (bottom right).

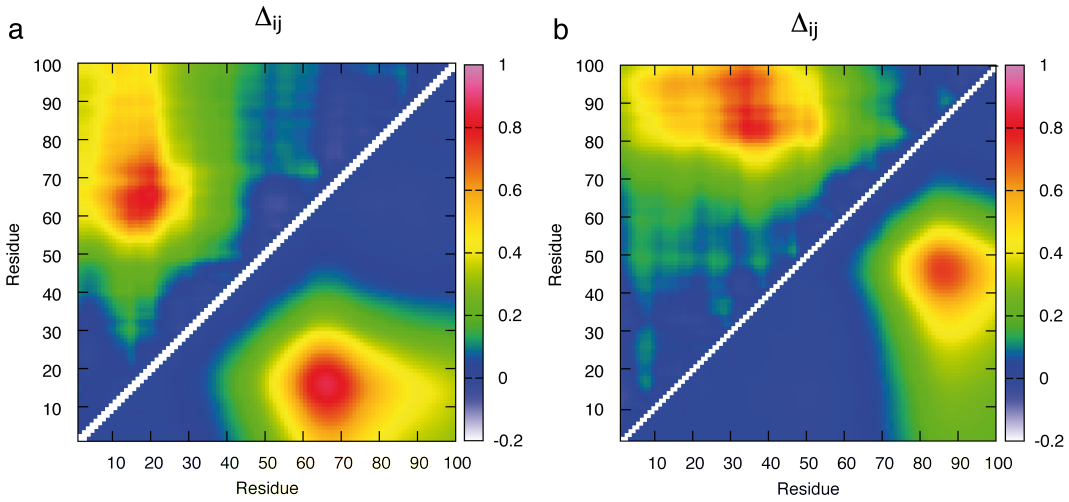

Figure 80 – Contact maps obtained with simulated data: (a) contact 11-20 and 61-70, (b) contact 41-50 and 81-90. ASTEROIDS ensembles (above the diagonal) and target ensembles (below the diagonal). The scale for the data above the diagonal has been multiplied by 0.50 for ease of contact identification.

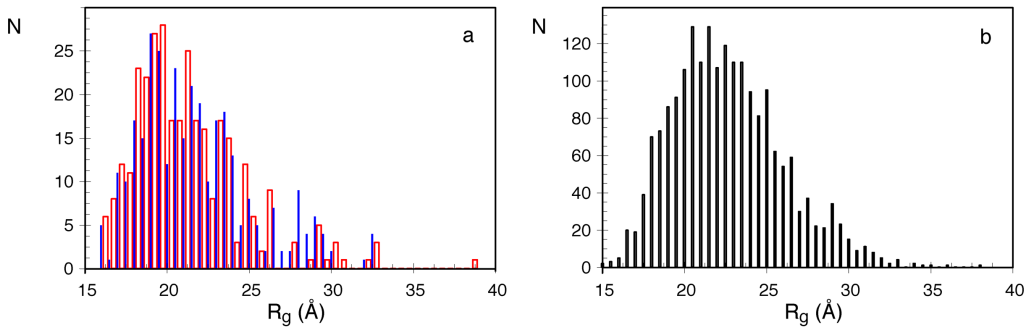

Figure 81 – Distribution of radii of gyration (R_g) obtained with simulated data involving a contact between 11-20 and 61-70. (a) Distribution of R_g in ASTEROIDS ensembles of size 80 (blue) or 160 (red) (b) Distribution of R_g for a set of 2,000 structures from the target ensembles (all structures have a contact between 11-20 and 61-70).

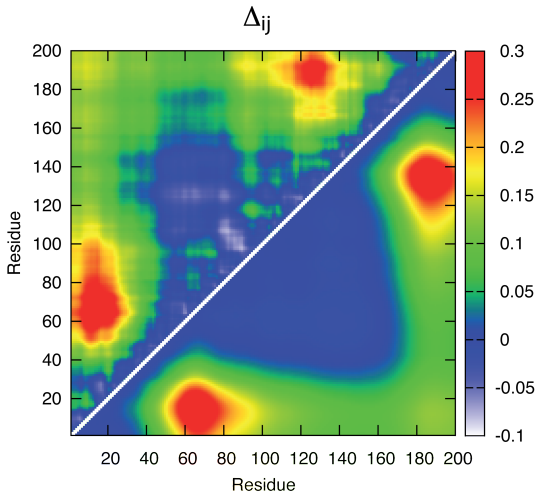

Figure 82 – Contact maps obtained with simulated data of two contacts between 11-20 and 61-70 or 41-50 and 81-90. ASTEROIDS ensembles (above the diagonal) and target ensembles (below the diagonal). The scale for the data above the diagonal has been multiplied by 0.66 for ease of contact identification.

underestimation of R_g compared to target ensembles and the increase of R_g with ensemble size have already been observed in several restrained molecular dynamics studies [140, 257, 284]. The FLEXIBLE-MECCANO-ASTEROIDS approach seems to suffer from similar features but less pronounced possibly because the dynamics of the spin label is explicitly taken into account or due to the inherent nature of the applied protocol. Here a sample and select approach is used where structures are selected from a pool of conformers, whereas the restrained MD approaches use direct refinement of the structures.

The capability of the ASTEROIDS approach to detect multiple contacts was tested on the 200 amino-acid protein simulated data. An identical protocol was applied and similar quality of data reproduction was obtained (data not shown). The resulting contact map is presented in Figure 82. The contact map allows a clear identification of the two contacts with similar level of accuracy as for a single contact, indicating that more complex systems with multiple contacts can be treated in a similar way.

11.3.2 *Application of the* ASTEROIDS *Approach to α-Synuclein Experimental PREs*

The approach was applied to α-Synuclein by selecting a sub-ensemble using ASTEROIDS among a pool of 10,000 FLEXIBLE-MECCANO conformers with no specific contacts. In order to determine the appropriate number of structures in the sub-ensemble and to test the validity of the invoked MTSL

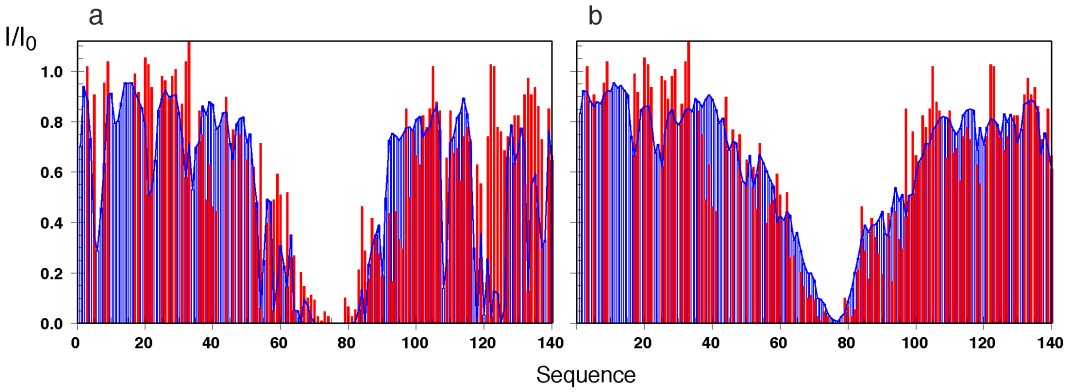

Figure 83 – Comparison of passive PRE data reproduction using static (a) and dynamic (b) models for the MTSL spin-label in α-Synuclein: (red) experimental data and (blue) ASTEROIDS selected ensemble of 200 structures.

side chain dynamics, the analysis was carried out for various ensemble sizes. The selection was done in a direct way including all four experimental paramagnetic probe data sets and in an indirect way where only three datasets corresponding to spin-labels at position 18, 90 and 140 were used. The data set corresponding to the spin label at position 76 was used for cross validation. Even if this indirect analysis is very demanding as a quarter of the data are removed, the quality of the obtained results justified the use of this protocol to determine optimal ensemble size. Interestingly, the model where the dynamics of the MTSL side chain is explicitly taken into account using the rotamer library reproduces the PRE data from the spin label at position 76 (passive data) better than a static description of the side chain (Figure 83). The corresponding RMSD is 0.17±0.01 for the dynamic model and 0.24±0.02 for the static description. The clear difference in data reproduction validates the use of a flexible side chain for the MTSL and in the following only this model will be considered.

The evolution of the radius of gyration, the indirect and direct data reproductions with the ensemble size can be found in Figure 84. All three parameters rapidly evolve for small numbers of conformers and reach a plateau between 100 and 200 structures. Therefore, an ensemble of 200 structures is considered as appropriate and the corresponding results in terms of direct data reproduction and contact maps are presented in Figure 85. Similar to previous studies, a clear contact appears between the C- and the N-terminal domains as well as a weaker interaction between the so-called NAC domain (residues 65-95) and the C-terminal domain [252, 256, 263].

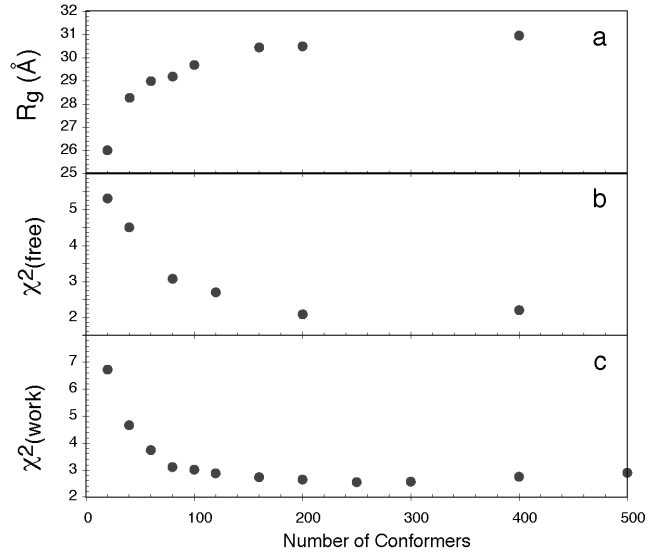

Figure 84 – Evolution of the radius of gyration (a), the indirect (b) and direct (c) data reproduction for α-Synuclein PREs as a function of the number of structures in the ensemble.

11.3.3 *Effect of Long-Range Order on RDCs in Unfolded Systems*

In order to determine the effect of long-range order on RDCs measured in unfolded proteins a set of Flexible-Meccano ensembles were generated. Each ensemble contained 100 000 members of a 100 amino-acid protein of arbitrary sequence. One of the ensembles had no specific contacts whereas the others had contacts between different parts of the peptide chain. $^1D_{NH}$ and $^1D_{C^\alpha H^\alpha}$ were calculated for each conformer and subsequently averaged over the whole ensemble. The alignment tensors were estimated using the full-length structures. The effects of the presence of contacts are presented in Figure 86. Looking at the overall shapes of the curves and the local variation of the RDCs, the presence of long-range order appears as a quenching of the RDC values of residues located between the interacting regions. Interestingly, the local conformational sampling of each amino-acid (data not shown) remains essentially unchanged in the presence of such RDC profile distortion. Those two points indicate that the possibility of using a combined LAW-baseline approach may be still valid if adequate baselines were found and that great care has to be taken when interpreting RDCs measured in unfolded proteins in terms of local conformational sampling only.

To further investigate the overall shapes of the RDC baselines in the presence of long-range order, similar ensembles of Flexible-Meccano conformers were generated but this time with a homo-polymer (poly-Valine). Results are

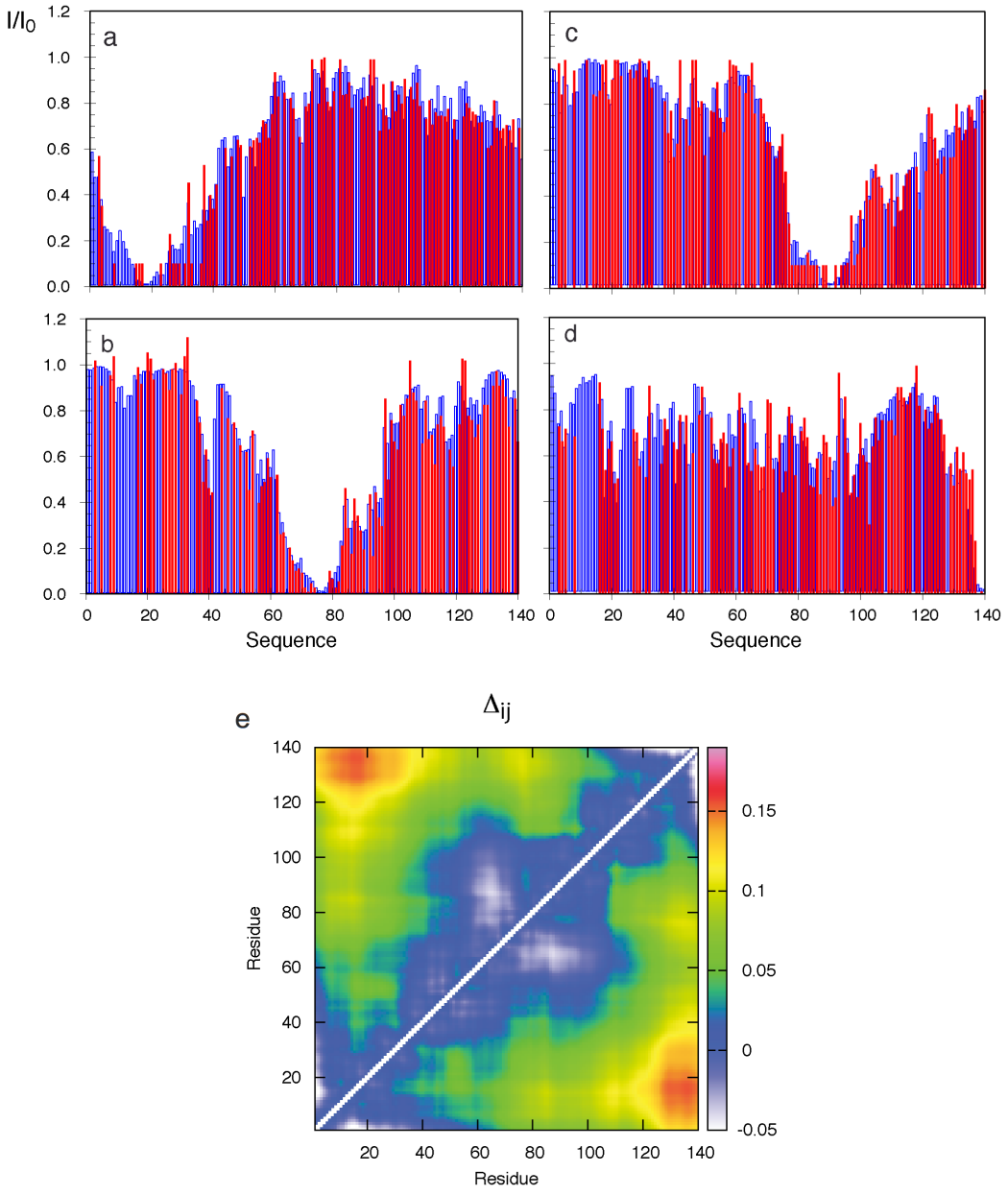

Figure 85 – Reproduction of PRE data measured for α-Synuclein (a-d) and corresponding contact map (e). (a-d) experimental data (red) and ASTEROIDS ensemble PRE reproduction (blue): (a) A18C, (b) A76C, (c) A90C and (d) A140C mutant data.

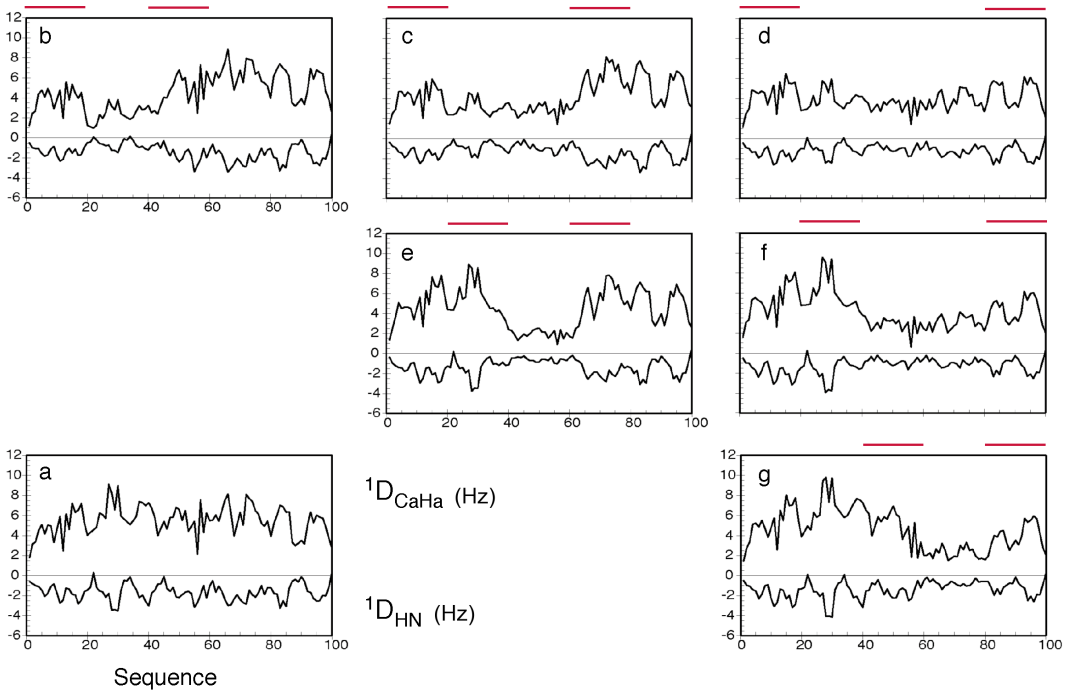

Figure 86 – Simulation of $^1D_{NH}$ and $^1D_{C^\alpha H^\alpha}$ RDC profiles for a disordered protein with an arbitrary sequence in the presence of contacts between different regions of the chain. (a) Profile of couplings in the absence of specific contacts. (b-g) Profiles of couplings in the presence of contacts between regions i and j: (b) i = 1-20, j = 41-60; (c) i = 1-20, j = 61-80; (d) i = 1-20, j = 81-100; (e) i = 21-40, j = 61-80; (f) i = 21-40, j = 81-100; (g) i = 41-60, j = 81-100. The continuous red bars above each plot indicate the positions of the contacting regions.

presented in Figure 87 with their corresponding parameterization according to the *ad hoc* equation 11.11.

As observed from the hetero-polymer curves, a clear quenching of the RDCs values of residues located between the two interacting regions can be observed.

As visible in Figure 87, the poly-Valine profiles can be well reproduced using a parameterization which consists of the combination of a hyperbolic cosine and Gaussian curves that ensure the quenching of RDCs between the two interacting regions. In order to test whether such a parameterization allows decorrelation of the local and global influence on RDCs calculations as in the absence of long-range contact, RDCs predicted using the LAW approach from a FLEXIBLE-MECCANO ensemble consisting of 200 conformers (with no specific contacts) were combined with the baseline containing a contact between 41-60 and 81-100. The resulting RDCs match closely those predicted from a 100,000-strong ensemble with global alignment tensor estimation where a contact was explicitly imposed between 41-60 and 81-100 (Figure 88). This description of the overall baseline in the presence

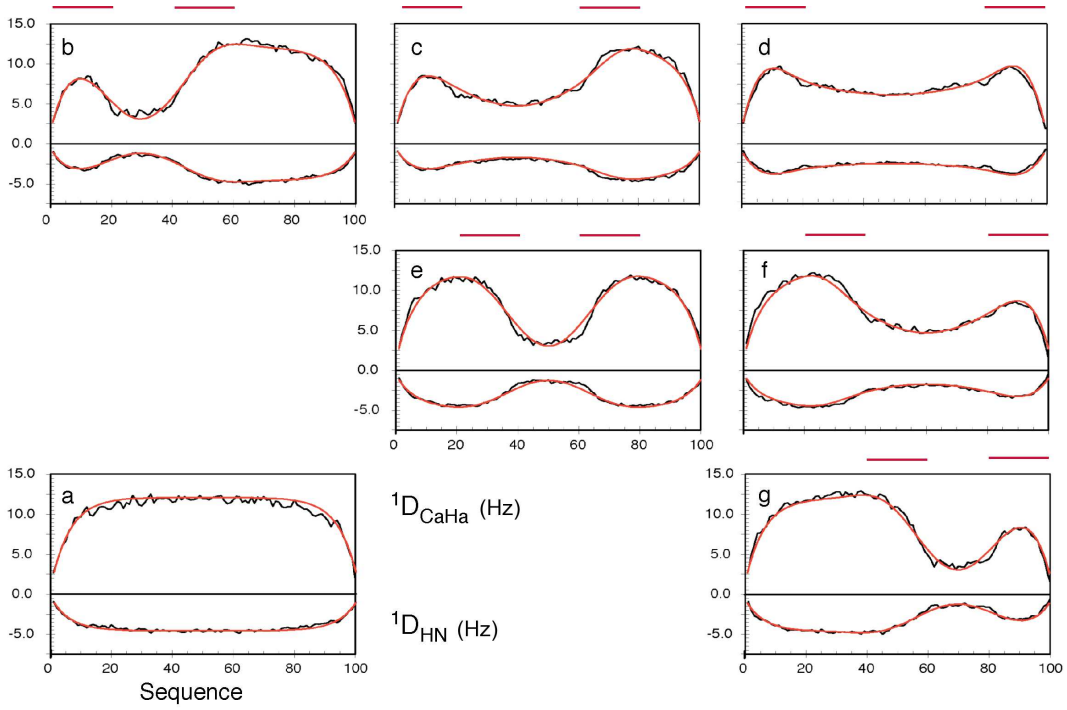

Figure 87 – Simulation of $^1D_{NH}$ and $^1D_{C^\alpha H^\alpha}$ RDC profiles for a poly-Valine homo-polymer in the presence of contacts between different regions (black) of the chain and their parameterization (red curves). (a) Profile of couplings in the absence of specific contacts. (b-g) Profiles of couplings in the presence of contacts between regions i and j: (b) i = 1-20, j = 41-60; (c) i = 1-20, j = 61-80; (d) i = 1-20, j = 81-100; (e) i = 21-40, j = 61-80; (f) i = 21-40, j = 81-100; (g) i = 41-60, j = 81-100. The continuous red bars above each plot indicate the positions of the contacting regions.

of long-range contacts can therefore be nicely combined with the local prediction of RDCs using the LAW approach.

11.3.4 *Combined Analysis of Simulated PRE and RDCs*

The combined LAW and baseline approach appears feasible even in the presence of long-range contacts which is particularly promising because the selection of large intractable ensembles of structures can be avoided. The approach was tested on the ensembles obtained in Section 11.3.1 on the basis of simulated PRE data on the 100 amino-acid protein of arbitrary sequence. The two contact maps were analyzed using the protocol described in Section 11.2.5 in order to determine the most populated contacts and the corresponding baselines were obtained using equation 11.11. The two baselines were combined with RDCs obtained by applying the LAW approach to the 200 conformers obtained in the PRE selection. The results and the comparison with the RDCs obtained from an average over 100,000 structures of

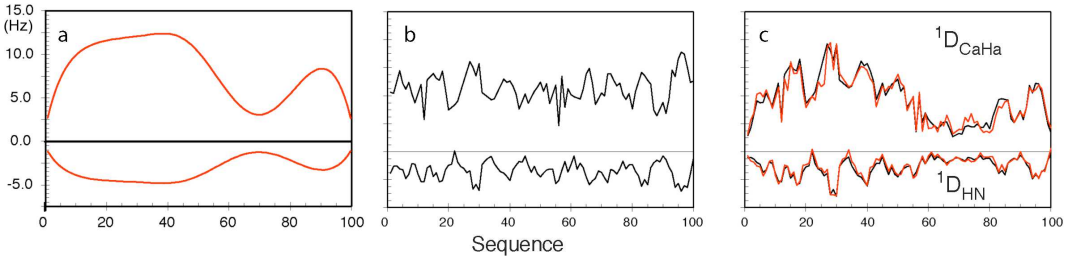

Figure 88 – Combination of baseline and RDCs averaged using the LAW approach. (a) Baseline calculated analytically for contacts between the regions centered on residues 50 and 90. (b) RDCs calculated using the LAW approach averaged over 200 structures. (c) Combination of the baseline from (a) and the local RDCs from (b) (red) compared to the full-length 100,000 conformers approach with explicit contact between 41-60 and 81-100 (black).

the appropriate target ensemble calculated using a global alignment tensor are shown in Figure 89.

The good agreement between the two approaches indicates that the separation of local and long-range order effects can be done accurately and that both PREs and RDCs can be combined in a meaningful way in ensembles of reasonable size.

11.3.5 *Combined Analysis of α-Synuclein PREs and RDCs*

A similar analysis was applied to α-Synuclein experimental PREs. The dominant long-range contact was extracted from the previously obtained contact map and the corresponding baseline was obtained using equation 11.11. The obtained ensemble of 200 structures was used to estimate RDCs with the LAW approach and the obtained results were compared to experimental RDCs measured in α-Synuclein aligned in PEG/hexanol (Figure 90).

The LAW-baseline approach provides a better data reproduction of experimental $^1D_{NH}$ RDCs (RMSD = 0.52) than a standard FLEXIBLE-MECCANO approach (RMSD = 0.72). This further validates the predicted effect of long-range order in RDCs profiles and shows that experimental PREs and RDCs can be combined in a meaningful way.

11.4 CONCLUSION

In order to describe the conformational behavior of IDPs, a molecular representation of the disordered state is required, based on diverse sources of structural data that often exhibit complex and very different averaging behavior. In this chapter we propose a combination of PREs and RDCs to

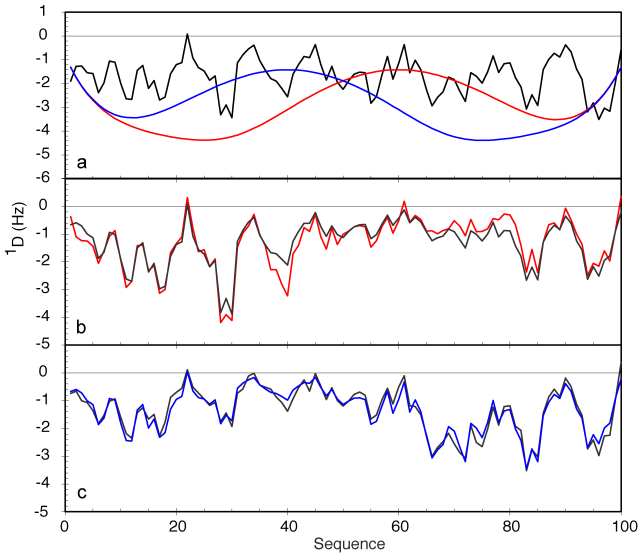

Figure 89 – Combined analysis of PREs and RDCs for simulated data. PREs were used to determine long-range contacts. (a) LAW-averaged RDCs (black), baseline extracted from the analysis of the target ensemble with contact between 11-20 and 61-70 (blue) or with contact between 41-50 and 81-90 (red). (b-c) comparison of RDCs obtained with explicit ensemble calculation using 100,000 conformers with appropriated contact (black) and RDCs obtained by combining LAW RDCs and baseline shown in (a) (colored curve). (b) contact between 41-50 and 81-90 (c) contact between regions 11-20 and 61-70.

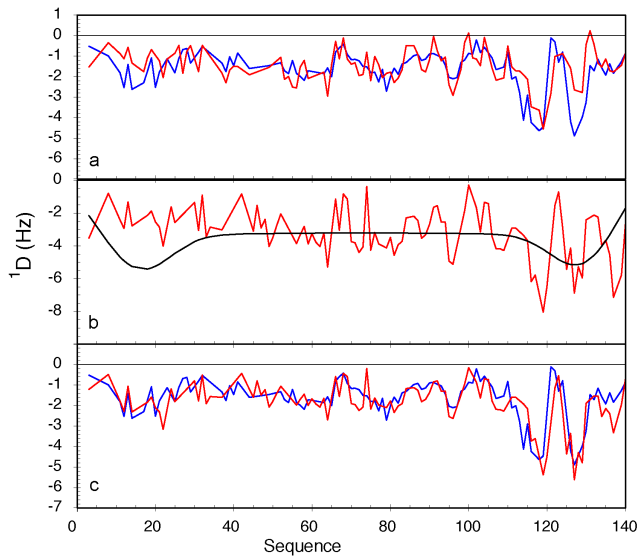

Figure 90 – Combined analysis of PREs and RDCs in α-Synuclein: comparison of experimental $^1D_{NH}$ RDCs with values obtained using the combination of LAW and baseline prediction from PRE analysis. (a) $^1D_{NH}$ RDCs experimental (blue) and calculated using a standard FLEXIBLE-MECCANO prediction (red). (b) LAW-predicted RDCs (red) and baseline derived from PRE analysis (black). (c) Combination of the curves shown in (b) (red) compared to the experimental $^1D_{NH}$ RDCs (blue).

study both long-range and local structural features of IDPs in solution. We demonstrate that ASTEROIDS, an ensemble selection algorithm, faithfully reproduces intramolecular contacts, even in the presence of highly diffuse, ill-defined target interactions, and show that explicit modeling of spin label mobility significantly improves reproduction of experimental PRE data, even in the case of highly disordered proteins.

Prediction of the effects of transient long-range contacts on RDC profiles reveals that weak intramolecular interaction can induce severe distortion of the profiles that compromises description of local conformational sampling if incorrectly accounted for. We develop an approach, based on PRE data to determine those long-range contacts and we parameterized their effects on the RDC profiles. This combined analysis is shown to be essential for the accurate interpretation of experimental data from α-Synuclein, confirming the presence of long-range order between distant regions in the protein.

In this study RDCs were described without taking to account the eventual correlation between long-range and the local properties. Of course this simple representation may not be sufficient in all experimental cases, e.g. if transient secondary structures are present in the system. Moreover the interference of local and global order effects may occur and in this case a more active refinement of the obtained ensemble against both PREs and RDCs is necessary.

Nevertheless, this study provides further insight into short-range and long-range effects in IDPs and takes a significant step towards a description of IDPs where all available experimental information can be combined to give an accurate and hopefully realistic description of such extremely flexible systems.

Part V

CONCLUSION

CONCLUSION

Understanding the biophysical principles and the biological function of protein flexibility remains a key challenge. The different studies presented in this Thesis aim, in diverse ways, to study this conformational disorder.

The first part of the Thesis focused on the dynamics of folded proteins. For these studies, RDCs provide powerful probes of molecular motions as they are sensitive to dynamics occurring on timescales up to the millisecond. One of their major strengths as conformational parameters resides in the possibility of deriving a complete analytical interpretation of experimental values. This possibility of developing, selecting and characterizing different biophysical models on the basis of RDCs relies on the opportunity to measure RDCs in different alignment media. As opposed to spin relaxation studies that are limited by the number of experimentally available data, RDC-based approaches can afford the luxury of using a model of molecular motion, assuming that the alignment medium does not interfere with the intrinsic dynamics.

Here the central model of the analytical RDC studies is the so-called GAF model. This model presents several advantages and, using diverse applications of the approach, is found to be well suited to the development of key insight into the presence of slow dynamics in proteins.

This model, which correspond to a Langevin oscillator, is expected to be able to accurately characterize peptide plane dynamics. Looking at dynamics as a stochastic diffusion process in a harmonic potential is a reasonable way to take into account thermal fluctuations in a structured system, approximating the potential in which angular diffusion occurs as a harmonic potential. The successful use of harmonic approximations in nearly all branches of Physics also justifies its use here — at least in a first approximation — and the quality of this approximation is tested and validated by comparing the results with model-free or molecular dynamics approaches.

This GAF model has been adapted here to quantitatively characterize, without recourse to other experimental data, the amount of dynamics presents in Ubiquitin. This study shows that RDCs alone are sufficient to accurately characterize proteins dynamics at timescales up to the millisecond, providing additional insight into the exploration of slow timescales dynamics. Interestingly this study revealed that slow motions are of small amplitude

in secondary structures, whereas loops and tails seem to exhibit significant motions on these timescales. One weakness of RDC-based studies is that no timescale information is available and therefore the distinction between slow and fast dynamics relies purely on the comparison with ^{15}N relaxation.

Part of this issue can be resolved using molecular dynamics based approaches. Slow motions can be expected to occur when a higher energy barrier in the potential energy landscape has to be overcome, which correspond to larger conformational changes. Various Ubiquitin ensemble descriptions, including the AMD ensemble presented here, have been now proposed and none indicate the existence of different low energies, well separated, solution clusters, which corresponds to the classical case of "very" slow dynamics. The picture provided by those approaches is more a nearly continuous ensemble of structures, populating a highly rough low energy region, with some excursions in more energetic conformations. This description is more coherent with a dynamic occurring an a broad range of timescales, than with a situation where fast and slow dynamics are clearly separated. Thus the distinction between those two — slow and fast — dynamics can appear as slightly artificial.

The AMD approach applied to Ubiquitin has brought a complementary view to the SF-GAF analysis. The AMD approach consists of a restraint-free molecular dynamics approach where conformational sampling is artificially enhanced in order to allow a characterization of the dynamics at timescales longer than those accessible through standard molecular dynamics. The correct level of acceleration is selected by estimating the quality of experimental data reproduction, which are here RDCs and J-couplings, both sensitive to timescales up to the millisecond. The comparison of the AMD and the SF-GAF approaches indicate a very good convergence in terms of order parameters. Even if this does not validate the two approaches from a purely logic point of view — both can be wrong and converge to the same results — considering the difference in hypothesis of the two approaches, their convergence can be seen as a very encouraging result that reinforces the confidence that can be gained from the two methods.

Nevertheless GAF models presents some drawbacks. One is the fact that all the motions are treated locally. Nonetheless this represents a significative advantage of the approach, as it allows the quantification of the amount of dynamics present in the system. The reinterpretation of these dynamics in terms of biophysical motions can lead to some undesired effects. For example, interpreting α- and β-motions in a structured system is problematic as it breaks the tetrahedral junctions if applied independently. Moreover internal motions are clearly not uncorrelated between planes. For example a previous GAF study of GB3 was able to identify correlations between planes

in the β-sheet using $^{3}J_{C'N}$ trans-hydrogen bond scalar couplings, whereas so-called anti-correlated crankshaft motions have been observed in various systems. More generally the presence of correlated motions is thought to be a way of propagating information and to be crucial for the slow dynamic description. Results obtained with approaches such as normal mode analysis, that treat motions as collective dynamics support these hypotheses. Large motions cannot occur simply as the sum of several completely independent motions, the presence of collectivity and correlation potentially increasing the rates and the specificity of such motions. From the NMR point of view the introduction of such collectivity may not be obvious as the measurements are site specific and therefore local analysis is immediately more intuitive. Nonetheless this site specificity provides *a priori* an extremely powerful source of information to characterize those motions.

Here an attempt to reinterpret RDCs dynamics using more elaborated `GAF` models has been proposed. The amount of dynamic presents in the system was estimated using a completely local description, namely the `SF-GAF` approach. Then the dynamics were reinvestigated in terms of both local and shared collective motions. In this context, no global motion can be detected by assuming in addition a local `3D-GAF` model of dynamics. Nonetheless modeling local dynamics with a γ-motion, which is the largest and the most physical `GAF` motion, allows the introduction of a shared `3D-GAF` motion to improve data reproduction. This approach, the newly introduced `3-1D-GAF`, led to very similar order parameter profiles compared to the standard `SF-GAF` approach, indicating that the motions invoked in this new approach remain reasonable. This analysis does not claim to give better data reproduction than the `3D-GAF` approach but simply to try to investigate how much of the dynamics measured in a particular site can be interpreted as a non-individual motion. Moreover the model developed here may be well adapted to characterize dynamics in complex systems, where for example a larger domain reorientations occur simultaneously with local dynamics within the domain, for example in large nucleic acid structures.

The approach developed here is a first attempt to characterize more precisely the nature of the motions occurring in folded proteins. It seems to be possible to interpret RDCs dynamics using approaches that are different to the completely uncorrelated local standard `3D-GAF`. It is worth noting that this kind of analysis explicitly requires a motional model and thus all RDC based model-free approaches are not well suited for such investigations. Even if the model proposed here can be reasonable for a β-sheet it is clearly not adapted to all kind of motion. Further investigations of how correlated and — or — domain motions can be reintroduced needs to be carried out. The most evident way to do it is to introduce inter-plane experimental

restraints such as long-range ^{1}H-^{1}H RDCs or $^{3}J_{C'N}$ trans-hydrogen bond scalar couplings. Ideally one would retain an analytical description, but the complexity of the model dramatically increases when considering interactions between nuclei in different planes. Therefore a numerical treatment may be more appropriate to investigate these questions. By using an SF-GAF in an initial first step the amount and directionality of local dynamics can be determined and in a second step the reinterpretation of the dynamics can be applied using the additional experimental information: for this last step the use of complementary approaches such as normal mode analysis or molecular dynamics may be useful.

The two following chapters focus on the characterization of a weak Ubiquitin SH3-C complex dynamics. The first step of the study aims to determined the fast and slow dynamics of the SH3-C, the Ubiquitin dynamics having already been investigated.

In the SH3-C study large RDC datasets have been collected in order to investigate the structural properties and slow dynamics of this system. The self-consistency of the obtained dataset was ensured by using a selection based on the SECONDA approach and the analysis of the data does not reveal inconsistencies by treating all datasets simultaneously. The data were analyzed in three different ways: with a completely static description using SCULPTOR, with a completely analytical GAF approach and with a sample and select method based on ensemble selection using ASTEROIDS-SVD. Concerning the structural properties the SCULPTOR and DYNAMIC-MECCANO approaches give very similar results, at apparently extremely high resolution compared to crystallographic structures of SH3-C domains. The DYNAMIC-MECCANO approach represents the third example of the use of RDCs alone to determine the structure and dynamics of the main chain of a small folded protein, with no aid from physical force fields or molecular dynamics based programs. The dynamic studies revealed slow dynamics in loop regions and N-terminal regions. Here GAF approaches were used and complemented with an approach that combined the selection efficiency of a genetic algorithm with the capability of SVD to obtain for a given conformational sampling the most adapted tensors. The nSrc loop exhibits high variability in the crystal structure and significant dynamics in solution, thus the question of how the conformation sampling in solution overlaps the crystal structure is an important one. Initial comparison suggests the presence of molecular recognition processes but more definitive answers require the comparison of the fast and slow conformational sampling of this loop in the free form and the complex. This cannot be achieved without an accurate study of the complex form of this SH3-C in interaction with its biological partner, Ubiquitin.

In the next step we have therefore investigated the complex dynamics using ^{15}N relaxation. This complex between Ubiquitin and SH3-C is very weak, and in this case it is impossible to find relevant experimental NMR conditions whereby the complex can be isolated from the free protein. We have therefore developed a titration approach that consists of measurement of relaxation in different mixtures of the proteins, to evaluate accurately the contribution from the complex. Linear evolution of the R_1 rates as a function of the fraction of the protein in the complex allowed a relatively straightforward extrapolation of the rates in the complex, as well as a refinement of the populations of the different mixtures of free and bound protein. R_2 rates present a more complex behavior as an exchange contribution has to be taken into account due to formation of the complex for all sites that do not present the same chemical environment in the complex and in the free protein. Using a model-free analysis of all of the different mixtures, correlated with an examination of the shifting resonances, the exchange contributions can be identified and used to determine the range of the kinetics of the complex formation. The relaxation rates were extrapolated for the complex after removal of the exchange contribution and analyzed. The analysis of the diffusion tensors showed a remarkably good convergence from the free protein tensors to diffusion tensor of the complex. Analyzing SH3-C and Ubiquitin alone or using an existing structure of the complex to analyze them simultaneously led to almost identical tensors. The structural information obtained from this study and previous RDCs based approach revealed very similar information content and thus the R_2/R_1 ratios that depend on the orientation of N_i-H_i^N vectors in the PAS of the tensor can be used to refine the complex structure. The local mobility was analyzed in terms of model-free analysis allowing to determine order parameters in the complex. The comparison with the free form fast dynamics leads to very similar results. N_i-H_i^N order parameter profiles are essentially similar with generally a slightly higher values in the complex. The only major difference being in the C-terminal part of the Ubiquitin: this tail, with its His-tag, being involved in the complex formation. However the study clearly reveals that, if drastic changes occurs on protein dynamics with the complex formation, this changes are mainly on slower timescales. In order to investigate this point, RDC based slow dynamic study has still to be achieved.

The second part of this Thesis is centered on the study of unfolded proteins. Unfolded systems do not behave as perfect random-coil homopolymers and deviation from such idealized descriptions of disordered chains represents an important challenge for structural biology and probably the basis of the biological function of these proteins. In order to study these highly flexible systems, the chosen approach is an ensemble selection procedure where the initial sampling is provided by Flexible-Meccano and the selection is achieved by a genetic algorithm Asteroids. The principle of this approach

is to generate a vast ensemble of structures representative of the random-coil state and to extract from it a sub-ensemble representative of the available experimental data. One of the interests of this approach is to obtain deviations from the random-coil without introducing hypothesis-driven direction or distortion of native amino-acid sampling. The approach supposes that the initial database provides a sufficiently large range of conformers to characterize the experimental system and the selection algorithm has to select the fraction of conformers that best reproduces experimental data within experimental error. Dealing with systems with extremely high number of degree of freedom, an extensive testing procedure was applied in order to investigate under which conditions the reproduction of experimental data leads to a meaningful representation of the protein conformational sampling.

This approach was used to characterize both local and global order in unfolded systems. This notion of order merits some brief discussion. The random-coil behavior consists of expected behavior in the presence of a perfectly unfolded proteins. This is an ideal situation that can be seen as the symmetric of a completely static description. For folded proteins dynamic conformational disorder can give further insight into the characterization of the system. Symmetrically for unfolded systems any deviation from the random-coil situation can be an important signature of the considered system, e.g. a tendency to sample more α-helix region can be used as a starting point for the description of the conformational equilibrium of transient secondary structures, that may play a role in the biophysical and biological understanding of the considered system.

Here the characterization of the local conformational behavior was analyzed in detail, providing methodological advances for the development of ensemble descriptions of unfolded proteins from extensive RDC measurements. Important observations were made concerning the parametric ranges that can deliver accurate conformational behavior. The detailed analysis of two experimental systems was undertaken. In the first case the conformational sampling of urea-denatured Ubiquitin was determined using RDCs. The obtained results allowed the identification of clear deviations from the random-coil situation. Particularly the sampling appeared to be more extended, possibly due to the presence of interacting urea with the protein backbone. Nevertheless no locally well defined secondary structures elements were identified. The second analysis was done for N_{TAIL} using chemical shifts. In this study, due to the high proportion of transient α-helices, the direct selection from the Flexible-Meccano conformers ensemble could not lead to satisfactory data reproduction. Thus a recursive refinement of the Flexible-Meccano (ϕ, ψ)-database was introduced during structure generation and selection. The obtained results clearly

indicated a tendency to form α-helix in the central part of the protein, as previously found with RDCs based studies. In this system the existence of the local order, translated into α-helix formation propensity, is clearly important for the interaction with biological partner as the region which is in direct interaction with its partner in the complex, folds into a complete helix upon binding.

The second part of unfolded protein study focused on the characterization of long-range order. This order, that can be expressed in terms of long-range contacts — or spatial proximity between different segments of the protein — was probed using PREs, highly sensitive parameters for the detection of distances between the paramagnetic probe and the studied nucleus. This allows the determination of the spatial proximity between the two terminal parts of the studied α-Synuclein. Further analysis revealed an important dependancy of RDCs on the presence of long-range contacts, that can be parameterized from the PREs detected contacts, and then used to ensure a correct interpretation of RDCs. This result demonstrates a new technique whereby local and global conformational information can be treated simultaneously in the same ensemble description.

The results obtained through this methodology are very promising and underline the possibility of using complementary data to characterize the unfolded state. Here RDCs, CSs and PREs were used separately in order to investigate different aspects of unfolded protein behavior and the information content of each quantity. CSs appeared as interesting probes of local conformational flexibility, PREs can efficiently determine long-range transient interactions and RDCs fill the gap between CSs and PREs as their are sensitive to both local and long-range order. Experimental restraints coming from other experimental technics could be incorporated too. Often NMR is complemented with Small Angle Scattering using both X-rays (SAXS) or neutrons (SANS). As they provide information about distance distribution functions, they provide a valuable complement to PREs. More generally any source of complementary information can be useful even if, as for example IR or Raman Spectroscopy, they mainly provide quantities averaged over all the peptide chain. Therefore a combined use of all those data can provide a much more accurate description of the unfolded state. Combining different sources of information is especially relevant for IDPs as due to their highly disordered behavior they require complex descriptions, e.g. large conformational ensembles, and thus are difficult to accurately and robustly characterize. Nevertheless, as demonstrated in these chapters, the highly diverse averaging properties of the different experimental parameters forces us to exercise extreme caution when combining the various sources of conformational characterization into the same ensemble description.

All the different studies presented here underline the complementary nature of conformational order and disorder. For folded protein the static structural description will provide important information, but accurate determination of the biophysical features requires the characterization of deviations from this ideal static situation. Studying the dynamic properties of folded system does not lead to a loss of this structural information and can even improve the static description by determining how the experimentally measured data are blurred by the dynamic disorder. As a demonstration of an important principle, the true dynamically averaged structures have been shown to improve the description of free experimental data for three different folded proteins, compared to single structures determined from dynamically averaged experimental restraints. If such an improvement can be measured on comparatively rigid and stable structures, one can expect that these approaches will have even greater significance in the case of proteins that experience broader conformational sampling.

If the intrinsic dynamic equilibrium is important for function, these conformational excursions must play a role in the formation of biological complexes. Even using rigid protein structures the kinetics of complex formation and dissociation impose a dynamic characterization of the system. Here too the study of the dynamics of this complex can lead to structural refinements considering the correlation obtained between RDCs measured in a steric alignment medium and the R_1/R_2 ratios underlying once more the intertwining of those two frequently opposed properties.

For unfolded system the random-coil situation can provide similar situations to the static description of folded system, as it consists in a limit, that allows one to understand the underlying nature of the studied systems but that has to be complemented by the reintroduction of specific conformational propensity. In fact to imagine interactions between two completely unfolded systems or their biological functions gives a similar level of understanding to what can be derived using a rigid structural approach for a folded system.

Therefore conformational order and disorder cannot be envisaged separately, with a full understanding of one aspect requiring the consideration of the other in order to be relevant.

The opposition between unfolded and folded states is not necessarily as important as expected, even if they are generally opposed and presented as two different sides of the (structural) biology. Both types of system can be studied using the same physical quantities, e.g. RDCs, spin relaxation rates... and often the methods and the models developed for the study of one phase can be applied to the other. The selection algorithm ASTEROIDS, developed for the unfolded state, here was almost directly applied, in

combination with SVD, to folded systems. Furthermore this approach may be a way to characterize partially folded system where both unfolded and structured motifs coexists. Similarly methods developed for folded system can be used to describe unfolded state dynamics. The GAF model may be applied to the unfolded state too. Direct application of purely GAF methods can be difficult, due to the necessity of large RDC datasets and the more complex way by which alignment tensors has to be treated. Nevertheless the inclusion of γ-motions in IDP ensemble representation may be relevant to complement the conformational flexibility description obtained using ensemble based description.

Even if the gap between folded and unfolded proteins remains important and even if the characterization of unfolded states remains insufficiently established to define a new, complete and coherent paradigm [285], the complementarity — more than the dichotomy — of both aspects starts to be clearly visible. This is why, I think, the nascent paradigm of "dynamically disordered biophysics" has not to be built in opposition to the structural biology paradigm, but should constitute a wider conceptual framework where the entanglement between conformational order and disorder can be fully expressed.

Part VI

ANNEXES

A

SINGULAR VALUE DECOMPOSITION

Singular Value Decomposition (SVD) is a method of factorization of rectangular matrices, than can be seen as an extension of diagonalisation for non-squared matrices.

Considering a $M \times N$ matrix $\mathbf{A}$ with $M \geqslant N$ whose coefficients are real numbers[1], it is possible to factorize $\mathbf{A}$ as [171]:

$$\mathbf{A} = \mathbf{U}\mathbf{W}\mathbf{V}^T \tag{A.1}$$

with:

- $\mathbf{U}$ a $M \times N$ column-orthogonal matrix.
- $\mathbf{W}$ a $N \times N$ diagonal matrix with positive or zero coefficients, which are by convention sorted by decreasing order of magnitude. Those coefficients are the singular values of $\mathbf{A}$ and their number corresponds to the rank r of A .
- $\mathbf{V}^T$ the transposed of $\mathbf{V}$ a $N \times N$ orthogonal matrix

The vectors of $\mathbf{U}$ that corresponds to non zero singular values generate a basis of the image of $\mathbf{A}$. The image of an ensemble of column vectors by $\mathbf{A}$ is all the possible vectors that can be generated by applying $\mathbf{A}$ on all the elements of the considered ensemble, i.e. all the possible $\mathbf{AX}$ values, $\mathbf{X}$ being in the starting ensemble.

The vectors of $\mathbf{V}$ that corresponds to zero singular values provide a basis for the kernel of $\mathbf{A}$. The kernel or nullspace of $\mathbf{A}$ being all the column vectors of the starting ensemble $\mathbf{X}$ fulfilling $\mathbf{AX} = 0$.

1 The exposed properties can be extended to matrices with complex coefficient too. Nevertheless only "real" matrices are used here, thus the complex case will not be considered any further.

SVD can be used for solving linear equation systems, which can be expressed in matricial form as:

$$\mathbf{A}\mathbf{X} = \mathbf{B} \tag{A.2}$$

where $\mathbf{X}$ is the unknown and $\mathbf{B}$ the targeted value, both expressed as a column vector.

If $\mathbf{B}$ is in the image of $\mathbf{A}$, the solution can be perfectly reached using SVD. Nevertheless if $\mathbf{B}$ does not lie in the image of $\mathbf{B}$, it is often of great interest to find the value of $\mathbf{X}$ that provides the closest solution to $\mathbf{B}$, i.e. to minimize [286, 287]:

$$\|\mathbf{A}\mathbf{X} - \mathbf{B}\|_2 \tag{A.3}$$

with $\|.\|_2$ the Euclidian norm defined as $\|X\|_2 = X^T X$.

The interest of SVD is to provide a pseudo-inverse[2] for $\mathbf{A}$, thus:

$$\mathbf{X} = \mathbf{V}\mathbf{W}^{-1}\mathbf{U}^T\mathbf{B} \tag{A.4}$$

with $\mathbf{W}^{-1}$ a matrix, whose coefficient are all zero except the first r diagonal coefficients that are the inverse of the singular values. The expression can be interpreted as follows:

1. $\mathbf{B}$ is projected on the image of $\mathbf{A}$, in $\mathbf{B}'$, which is the closest solution in the sense of the Euclidian norm [288].

2. the inverse transformation is done in order to obtain one antecedent of $\mathbf{B}'$, $\mathbf{X}$

Nevertheless if the $r < N$ the antecedents of $\mathbf{B}'$ are not unique as any vector of the kernel can be added to $\mathbf{X}$ and lead to the same image $\mathbf{B}'$. As previously mentioned, the $\mathbf{V}$ matrix is build to have a basis of the kernel defined with the vectors that correspond to zero singular values. Thus using SVD, the obtained antecedent is the one with the smallest norm, i.e. with no component in the kernel, which gives a condition to ensure a pseudo-unicity of the results [287].

2 $\mathbf{A}$ cannot be invertible if $M > N$ or $M = N$ and the rank of $\mathbf{A}$ is smaller than N.

This approach can be usefully applied to RDC analysis [148, 289] in order to determined optimal tensors for a given structure or an ensemble of (aligned) structures, using:

$$\boldsymbol{A} = \begin{pmatrix} \left\langle \overline{x}_1^2 - \overline{y}_1^2 \right\rangle & \left\langle \overline{x}_1^2 - \overline{y}_1^2 \right\rangle & \left\langle 2\overline{x}_1\overline{y}_1 \right\rangle & \left\langle 2\overline{x}_1\overline{z}_1 \right\rangle & \left\langle 2\overline{y}_1\overline{z}_1 \right\rangle \\ \left\langle \overline{x}_2^2 - \overline{y}_2^2 \right\rangle & \left\langle \overline{x}_2^2 - \overline{y}_2^2 \right\rangle & \left\langle 2\overline{x}_2\overline{y}_2 \right\rangle & \left\langle 2\overline{x}_2\overline{z}_2 \right\rangle & \left\langle 2\overline{y}_2\overline{z}_2 \right\rangle \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \left\langle \overline{x}_N^2 - \overline{y}_N^2 \right\rangle & \left\langle \overline{x}_N^2 - \overline{y}_N^2 \right\rangle & \left\langle 2\overline{x}_N\overline{y}_N \right\rangle & \left\langle 2\overline{x}_N\overline{z}_N \right\rangle & \left\langle 2\overline{y}_N\overline{z}_N \right\rangle \end{pmatrix} \tag{A.5}$$

$$X = \begin{pmatrix} A_{yy} \\ A_{zz} \\ A_{xy} \\ A_{xz} \\ A_{yz} \end{pmatrix} \qquad \text{and} \qquad B = \begin{pmatrix} D_1^{red} \\ D_2^{red} \\ \vdots \\ D_N^{red} \end{pmatrix} \tag{A.6}$$

where N is the total number of RDCs measured in an alignment medium, $(\overline{x}_i, \overline{y}_i, \overline{z}_i)$ the normalized component of the i-th internuclear vector of interest, $\langle . \rangle$ the average over all the used structures, A_{jj} the five independent elements of the Saupe matrix and D_i^{red} the i-th experimental RDCs normalized by the corresponding d_{is}. The matricial representation is derived directly from the cosine director description in Section 2.4 and different kinds of couplings, e.g. ${}^1D_{NH}$, ${}^1D_{C'C^\alpha}$ ... can be used simultaneously. The SVD of $\boldsymbol{A}$ will provide a way to estimate the most appropriate tensor ($\mathbf{X}$).

Thus, using SVD it is possible to obtain simultaneously all the optimal data reproduction and the optimal tensors for any given ensemble of aligned structures and a representation of the tensors using axial and rhombic components and Euler angles, can easily be obtained by diagonalization of the reconstituted full Saupe matrix.

B

DRAMATIS PERSONAE

UBIQUITIN is an important protein, involved in extremely divers cellular processes, including endocytosis, cell-cycle control, DNA repair, gene regulation and expression [8, 290, 291]. Its most famous function is ubiquitination, that consists in tagging poly-ubiquitin chains and protein to induce their degradation in the proteasome. A large variety of poly-ubiquitin exists inducing different reaction pathways [292].

CD2AP SH3-C. SH3 domains are very common molecular recognition elements [293]. Even though many SH3 domains exist, with different sequence compositions, they often exhibit high structural similarity [294].The CD2 associated protein (CD2AP) is involved in kidney function [203] and the considered CD2AP SH3-C is a part of a recognition complex made up of three SH3 domains A, B and C [295] that all interact with Ubiquitin.

PROTEIN GB3 is the B3 domain of an Ig-binding domain of streptococcal protein G and is involved in antibody-antigen recognition by interacting with its physiological partner, the antigen-binding domain of IgG [8, 164].

N_{TAIL} is a partially unfolded protein. It is the C-terminal domain of the nucleoprotein of Sendai virus, that causes bronchiolitis in mice and primates [296]. N_{TAIL} interacts with the C-terminal part of a phosphoprotein P_X (a partially folded protein) and thereby initiates the viral RNA transcription and replication [297].

α-SYNUCLEIN is an IDP of 140 amino-acids presents in neural tissues and which has an important role in the development of neuro-degenerative diseases, especially Parkinson disease [298]. The apparition of the so-called Lewy bodies are related to the disease [299]. They consist of abnormal aggregates of protein, which are visible under the microscope. α-Synuclein is the main constituent of these bodies [300, 301].

C

TABLES AND SUPPORTING FIGURES

C.1 NUMERICAL VALUES TABLES AND ADDITIONAL FIGURES FOR UBIQUITIN SF-GAF ANALYSIS

Table 17 – Ubiquitin order parameters from SF-GAF analysis. Order parameters for N_i-H_i^N, C'_{i-1}-N_i, C^{α}_{i-1}-C'_{i-1} and C'_{i-1}-H_i^N vector for each studied peptide plane.

Plane	S^2_{NH}	S^2_{CN}	S^2_{CC}	S^2_{CH}
2	0.84 ± 0.02	0.84 ± 0.02	0.83 ± 0.02	0.84 ± 0.02
3	0.89 ± 0.03	0.86 ± 0.03	0.83 ± 0.04	0.89 ± 0.03
4	0.89 ± 0.02	0.93 ± 0.01	0.96 ± 0.00	0.89 ± 0.02
5	0.81 ± 0.03	0.79 ± 0.01	0.77 ± 0.00	0.81 ± 0.03
6	0.92 ± 0.02	0.95 ± 0.01	0.97 ± 0.00	0.92 ± 0.02
7	0.73 ± 0.02	0.79 ± 0.02	0.84 ± 0.02	0.74 ± 0.02
8	0.57 ± 0.03	0.75 ± 0.02	0.91 ± 0.01	0.60 ± 0.03
11	0.48 ± 0.02	0.45 ± 0.03	0.42 ± 0.05	0.47 ± 0.02
12	0.70 ± 0.02	0.80 ± 0.02	0.88 ± 0.02	0.71 ± 0.02
13	0.72 ± 0.03	0.81 ± 0.03	0.88 ± 0.03	0.73 ± 0.03
14	0.88 ± 0.03	0.83 ± 0.04	0.80 ± 0.06	0.87 ± 0.03
15	0.90 ± 0.04	0.85 ± 0.04	0.82 ± 0.05	0.89 ± 0.03
16	0.93 ± 0.01	0.92 ± 0.01	0.91 ± 0.00	0.92 ± 0.01
17	0.89 ± 0.02	0.90 ± 0.01	0.91 ± 0.00	0.89 ± 0.02
18	0.73 ± 0.01	0.82 ± 0.01	0.89 ± 0.00	0.74 ± 0.01
20	0.74 ± 0.04	0.85 ± 0.02	0.94 ± 0.01	0.75 ± 0.04
21	0.86 ± 0.03	0.85 ± 0.03	0.85 ± 0.03	0.86 ± 0.03
22	0.93 ± 0.05	0.92 ± 0.02	0.91 ± 0.01	0.92 ± 0.04
23	0.88 ± 0.02	0.87 ± 0.02	0.87 ± 0.02	0.88 ± 0.02
25	0.89 ± 0.02	0.82 ± 0.01	0.78 ± 0.00	0.88 ± 0.01
27	0.86 ± 0.05	0.89 ± 0.04	0.90 ± 0.04	0.87 ± 0.05

Continued on next page

Plane	S^2_{NH}	S^2_{CN}	S^2_{CC}	S^2_{CH}
28	0.96 ± 0.02	0.95 ± 0.01	0.95 ± 0.01	0.96 ± 0.02
29	0.86 ± 0.01	0.92 ± 0.01	0.96 ± 0.00	0.87 ± 0.01
30	0.85 ± 0.02	0.88 ± 0.01	0.90 ± 0.00	0.85 ± 0.02
31	0.93 ± 0.02	0.92 ± 0.01	0.91 ± 0.01	0.92 ± 0.02
32	0.95 ± 0.02	0.97 ± 0.01	0.98 ± 0.01	0.96 ± 0.02
33	0.85 ± 0.03	0.88 ± 0.04	0.90 ± 0.05	0.85 ± 0.03
34	0.85 ± 0.02	0.88 ± 0.01	0.90 ± 0.00	0.85 ± 0.02
35	0.81 ± 0.03	0.80 ± 0.03	0.80 ± 0.03	0.81 ± 0.03
36	0.70 ± 0.01	0.72 ± 0.05	0.74 ± 0.08	0.70 ± 0.01
39	0.87 ± 0.06	0.91 ± 0.03	0.94 ± 0.01	0.88 ± 0.06
40	0.72 ± 0.03	0.84 ± 0.02	0.94 ± 0.01	0.73 ± 0.03
41	0.64 ± 0.02	0.76 ± 0.01	0.87 ± 0.01	0.65 ± 0.02
42	0.84 ± 0.03	0.77 ± 0.03	0.72 ± 0.04	0.83 ± 0.03
43	0.85 ± 0.03	0.88 ± 0.01	0.90 ± 0.01	0.85 ± 0.02
44	0.89 ± 0.01	0.92 ± 0.01	0.94 ± 0.00	0.90 ± 0.01
45	0.81 ± 0.01	0.86 ± 0.01	0.90 ± 0.00	0.81 ± 0.01
47	0.76 ± 0.02	0.83 ± 0.02	0.89 ± 0.02	0.77 ± 0.02
48	0.68 ± 0.02	0.73 ± 0.02	0.76 ± 0.02	0.69 ± 0.02
49	0.84 ± 0.03	0.88 ± 0.03	0.90 ± 0.03	0.84 ± 0.03
50	0.72 ± 0.02	0.81 ± 0.02	0.89 ± 0.01	0.74 ± 0.02
51	0.68 ± 0.02	0.75 ± 0.03	0.80 ± 0.03	0.69 ± 0.02
52	0.67 ± 0.01	0.62 ± 0.05	0.58 ± 0.07	0.67 ± 0.01
54	0.82 ± 0.03	0.87 ± 0.02	0.90 ± 0.02	0.83 ± 0.03
55	0.84 ± 0.01	0.91 ± 0.01	0.96 ± 0.00	0.85 ± 0.01
56	0.84 ± 0.02	0.87 ± 0.01	0.90 ± 0.00	0.84 ± 0.02
57	0.84 ± 0.03	0.91 ± 0.02	0.96 ± 0.00	0.85 ± 0.03
58	0.95 ± 0.03	0.97 ± 0.01	0.98 ± 0.01	0.96 ± 0.03
59	0.92 ± 0.02	0.95 ± 0.01	0.97 ± 0.00	0.92 ± 0.02
60	0.66 ± 0.02	0.76 ± 0.01	0.84 ± 0.01	0.67 ± 0.02
61	0.85 ± 0.03	0.88 ± 0.03	0.90 ± 0.03	0.85 ± 0.03
62	0.51 ± 0.01	0.70 ± 0.01	0.90 ± 0.00	0.53 ± 0.01
63	0.83 ± 0.02	0.91 ± 0.01	0.96 ± 0.00	0.84 ± 0.02
64	0.94 ± 0.02	0.68 ± 0.03	0.55 ± 0.04	0.90 ± 0.02
65	0.61 ± 0.03	0.53 ± 0.04	0.48 ± 0.06	0.60 ± 0.03
66	0.83 ± 0.03	0.90 ± 0.02	0.96 ± 0.01	0.84 ± 0.03

Continued on next page

Plane	S^2_{NH}	S^2_{CN}	S^2_{CC}	S^2_{CH}
67	0.87 ± 0.03	0.89 ± 0.01	0.91 ± 0.00	0.87 ± 0.03
68	0.90 ± 0.01	0.93 ± 0.01	0.94 ± 0.00	0.91 ± 0.01
70	0.79 ± 0.03	0.85 ± 0.02	0.89 ± 0.01	0.80 ± 0.03
71	0.54 ± 0.03	0.69 ± 0.02	0.82 ± 0.02	0.56 ± 0.03
72	0.58 ± 0.03	0.65 ± 0.05	0.71 ± 0.08	0.59 ± 0.03
74	0.29 ± 0.03	0.34 ± 0.07	0.38 ± 0.15	0.30 ± 0.03
76	0.01 ± 0.01	0.08 ± 0.02	0.14 ± 0.06	0.02 ± 0.01

Table 18 – Ubiquitin order parameters from SF-GAF analysis. Order parameters for N_i-H^N_i, C'_{i-1}-N_i, C^α_{i-1}-C'_{i-1} and C'_{i-1}-H^N_i vector for each studied peptide plane.

Plane	σ_α	σ_β	σ_γ
2	4.26 ± 8.56	13.73 ± 0.89	0.00 ± 1.73
3	14.70 ± 2.72	0.00 ± 1.37	11.18 ± 2.20
4	4.26 ± 7.52	3.68 ± 3.02	11.32 ± 1.60
5	15.79 ± 4.62	6.65 ± 2.11	13.64 ± 1.56
6	4.26 ± 7.75	2.63 ± 2.39	9.46 ± 1.25
7	4.26 ± 5.66	12.53 ± 1.13	13.85 ± 0.62
8	4.26 ± 6.78	2.22 ± 3.39	26.83 ± 1.49
11	26.54 ± 3.48	17.02 ± 0.46	22.56 ± 1.44
12	4.26 ± 6.96	9.24 ± 1.43	18.32 ± 0.86
13	4.26 ± 5.20	9.24 ± 1.57	17.21 ± 1.22
14	16.27 ± 3.35	0.00 ± 1.34	11.89 ± 2.08
15	14.00 ± 2.50	5.61 ± 3.45	9.19 ± 1.11
16	4.26 ± 5.28	9.24 ± 2.70	0.00 ± 2.14
17	4.26 ± 6.08	9.24 ± 3.28	6.85 ± 2.29
18	4.26 ± 6.96	9.24 ± 2.25	16.65 ± 0.54
20	4.26 ± 4.73	3.72 ± 3.47	18.85 ± 1.82
21	4.26 ± 7.32	12.90 ± 1.56	2.11 ± 2.96
22	4.26 ±10.20	9.24 ± 4.53	0.00 ± 4.83
23	4.26 ± 9.47	12.03 ± 1.06	0.00 ± 1.34
25	17.04 ± 5.38	3.66 ± 2.91	10.55 ± 1.02
27	4.26 ± 4.32	9.24 ± 2.59	8.91 ± 1.48
28	4.26 ± 4.50	6.56 ± 0.70	0.00 ± 0.99

Continued on next page

Plane	σ_α	σ_β	σ_γ
29	4.26 ± 9.62	2.51 ± 3.03	13.16 ± 2.55
30	4.26 ± 8.47	9.24 ± 1.10	9.90 ± 1.06
31	4.26 ±11.43	9.24 ± 0.90	0.00 ± 1.55
32	4.26 ± 8.23	0.00 ± 0.92	7.32 ± 2.23
33	4.26 ± 3.05	9.24 ± 1.49	10.01 ± 0.93
34	4.26 ± 8.26	9.24 ± 1.73	10.07 ± 1.52
35	4.26 ± 6.74	15.47 ± 1.46	0.00 ± 2.94
36	18.44 ± 4.75	0.00 ± 0.50	20.65 ± 0.90
39	4.26 ± 7.76	5.53 ± 3.18	11.30 ± 4.31
40	4.26 ± 5.56	2.69 ± 3.34	20.01 ± 1.39
41	4.26 ± 5.73	9.24 ± 1.10	21.33 ± 1.32
42	18.52 ± 2.66	7.35 ± 3.14	11.21 ± 0.99
43	4.26 ± 4.09	9.24 ± 0.81	9.89 ± 2.21
44	4.26 ± 6.28	6.28 ± 1.93	9.37 ± 0.83
45	4.26 ± 8.62	9.24 ± 2.74	12.51 ± 0.86
47	4.26 ± 4.74	9.24 ± 1.01	15.34 ± 0.75
48	4.26 ± 6.37	16.30 ± 1.29	13.06 ± 2.32
49	4.26 ± 8.59	9.24 ± 1.80	10.54 ± 2.18
50	4.26 ± 9.90	9.24 ± 0.82	16.97 ± 1.08
51	4.26 ± 6.81	14.33 ± 1.66	15.14 ± 1.24
52	19.52 ± 4.12	15.55 ± 1.18	13.35 ± 1.77
54	4.26 ± 7.87	9.24 ± 1.74	11.57 ± 2.36
55	4.26 ± 6.41	0.00 ± 2.67	14.44 ± 0.77
56	4.26 ± 7.10	9.24 ± 3.63	10.68 ± 1.58
57	4.26 ±10.69	0.00 ± 3.47	14.62 ± 1.63
58	4.26 ± 7.35	0.00 ± 1.39	7.26 ± 3.00
59	4.26 ± 5.95	0.00 ± 3.29	10.15 ± 1.31
60	4.26 ± 5.12	11.75 ± 0.80	18.55 ± 1.58
61	4.26 ± 5.44	9.24 ± 1.82	10.07 ± 1.24
62	4.26 ± 6.18	0.00 ± 3.38	30.53 ± 0.77
63	4.26 ± 9.82	0.00 ± 3.55	14.74 ± 1.20
64	28.85 ± 2.23	4.56 ± 3.08	0.00 ± 0.93
65	30.36 ± 3.80	5.68 ± 4.10	22.82 ± 2.36
66	4.26 ± 6.45	0.00 ± 1.98	14.95 ± 1.91
67	4.26 ± 8.81	9.24 ± 3.39	8.65 ± 2.67

Continued on next page

Plane	σ_α	σ_β	σ_γ
68	4.26 ± 6.13	6.12 ± 1.48	8.92 ± 1.10
70	4.26 ± 5.19	9.24 ± 1.11	13.64 ± 2.19
71	4.26 ± 7.25	11.52 ± 1.14	24.85 ± 1.51
72	4.26 ±17.45	18.35 ± 5.29	16.51 ± 4.16
74	4.26 ±20.95	34.92 ± 9.15	18.19 ± 9.25
76	4.26 ±26.74	49.61 ± 6.64	62.29 ± 4.72

C.2 COMPARISON OF ORDER PARAMETERS FROM ^{15}N RELAXATION MEASUREMENTS

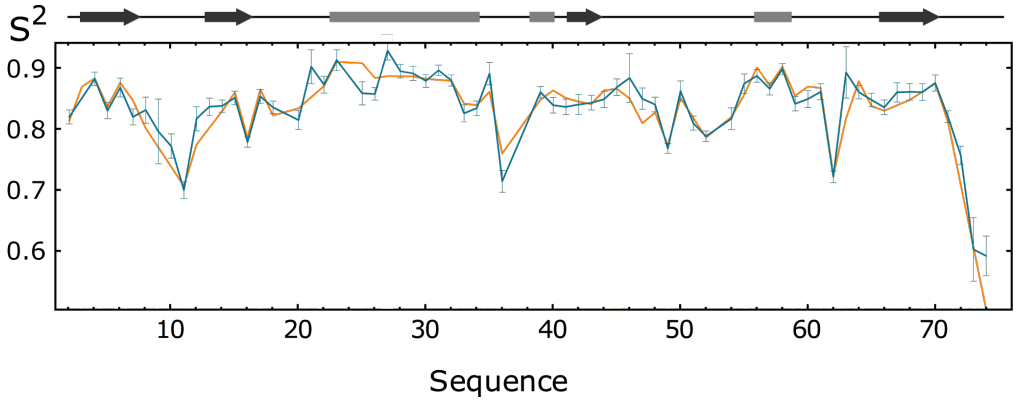

Figure 91 – Comparison between two independently measured ^{15}N relaxation N_i-H_i^N order parameters in Ubiquitin: (orange) ^{15}N relaxation derived order parameters from Lienin et al. [149] (turquoise) ^{15}N relaxation derived from the relaxation data obtained in the Chapter 8. Grey boxes indicate α-helix and darker arrows indicate β-sheet.

C.3 STRUCTURE-FREE ANALYSIS OF UBIQUITIN USING 1.024A BOND LENGTH

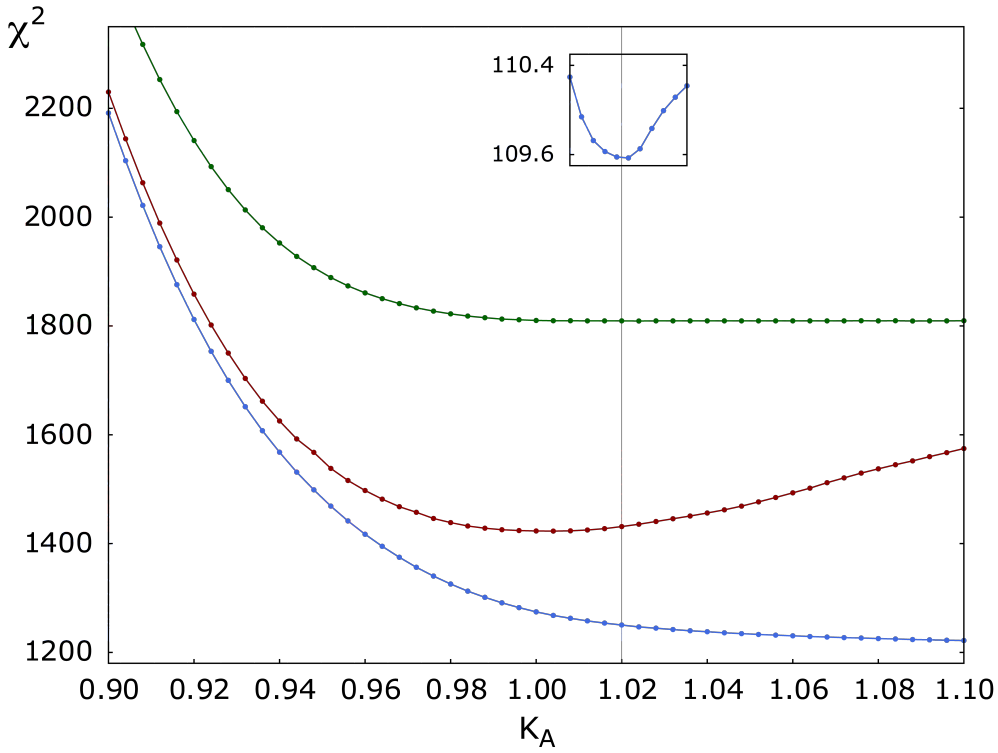

Figure 92 – Ubiquitin experimental data. Effect of the alignment tensor scaling on direct data reproduction χ^2 according to model `S` (green), `1D-GAF` (red), `3D-GAF` (blue), using an N_i-H_i^N length of 1.024 Å. The scaling is done according to K_A. The value $K_A = 1$ corresponds to the tensors obtained after `1D-GAF` optimization. The inset corresponds to indirect data reproduction according to `3D-GAF`. Gray line corresponds to the optimal tensors.

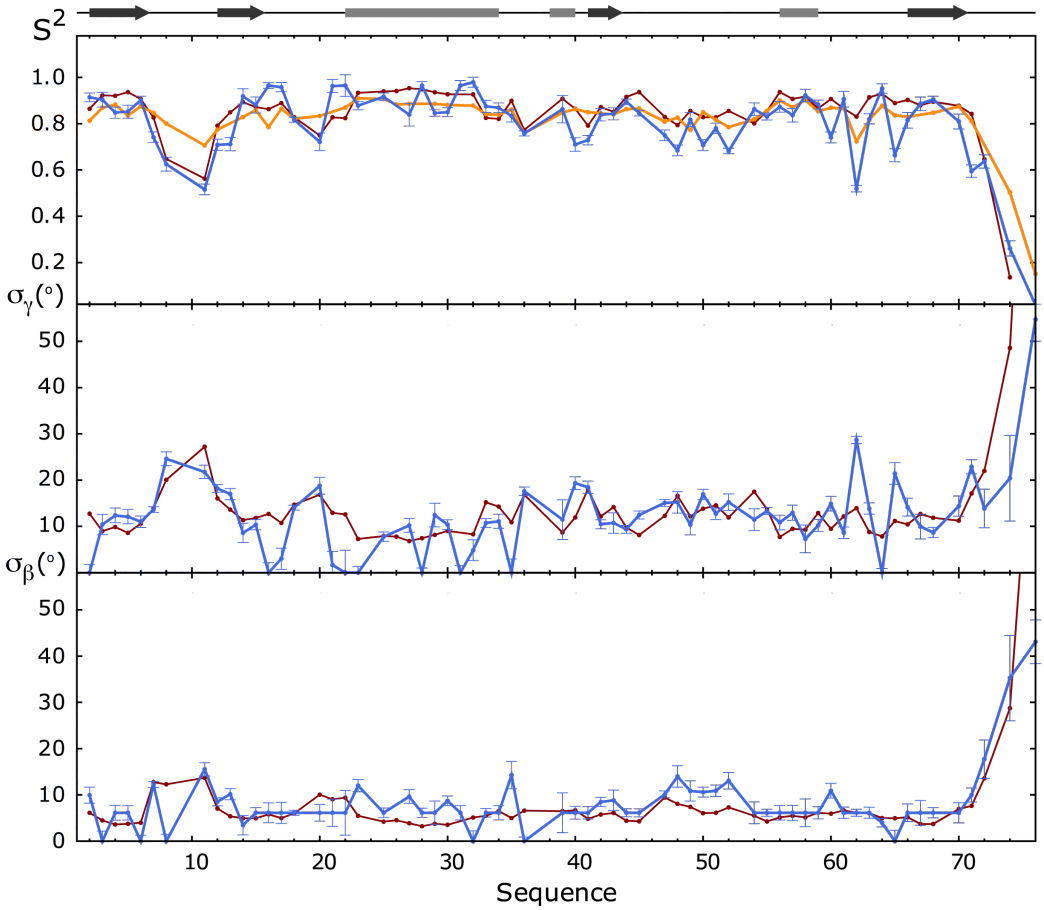

Figure 93 – Local Ubiquitin dynamics obtained through SF-GAF analysis, using an N_i-H_i^N length of 1.024 Å. N_i-H_i^N order parameters (upper panel) and amplitudes of reorientations for γ-motion (central panel) and β-motion (lower panel) derived from SF-GAF analysis (blue), 40ons MD simulation (red) or ^{15}N relaxation (orange). Secondary structures are indicated on the top of the figure. Grey boxes indicate α-helix and darker arrows indicate β-sheet.

C.4 NUMERICAL VALUES TABLES FOR GB3 SF-GAF ANALYSIS

Table 19 – GB3 order parameters from SF-GAF analysis. Order parameters for N_i-H_i^N, C'_{i-1}-N_i, C^{α}_{i-1}-C'_{i-1} and C'_{i-1}-H_i^N vector for each studied peptide plane.

Plane	S^2_{NH}	S^2_{CN}	S^2_{CC}	S^2_{CH}
8	0.76 ± 0.04	0.83 ± 0.02	0.88 ± 0.01	0.77 ± 0.04
9	0.82 ± 0.06	0.86 ± 0.03	0.89 ± 0.01	0.83 ± 0.05
10	0.91 ± 0.02	0.90 ± 0.01	0.90 ± 0.00	0.91 ± 0.02
11	0.81 ± 0.06	0.85 ± 0.03	0.88 ± 0.01	0.81 ± 0.05
12	0.72 ± 0.04	0.80 ± 0.02	0.87 ± 0.01	0.73 ± 0.03
13	0.91 ± 0.05	0.90 ± 0.02	0.90 ± 0.01	0.91 ± 0.04

Continued on next page

Plane	S^2_{NH}	S^2_{CN}	S^2_{CC}	S^2_{CH}
14	0.84 ± 0.01	0.83 ± 0.01	0.82 ± 0.01	0.83 ± 0.01
15	0.66 ± 0.06	0.75 ± 0.04	0.83 ± 0.03	0.67 ± 0.06
16	0.67 ± 0.04	0.72 ± 0.03	0.77 ± 0.04	0.68 ± 0.04
17	0.47 ± 0.03	0.65 ± 0.03	0.82 ± 0.01	0.49 ± 0.03
18	0.76 ± 0.03	0.86 ± 0.02	0.95 ± 0.02	0.78 ± 0.03
19	0.60 ± 0.06	0.73 ± 0.04	0.85 ± 0.01	0.61 ± 0.06
20	0.72 ± 0.08	0.81 ± 0.04	0.87 ± 0.01	0.74 ± 0.07
21	0.81 ± 0.03	0.76 ± 0.03	0.72 ± 0.04	0.81 ± 0.03
22	0.88 ± 0.09	0.89 ± 0.05	0.89 ± 0.01	0.88 ± 0.08
23	0.81 ± 0.02	0.85 ± 0.01	0.88 ± 0.00	0.81 ± 0.02
24	0.71 ± 0.03	0.80 ± 0.02	0.87 ± 0.01	0.72 ± 0.03
25	0.74 ± 0.02	0.81 ± 0.01	0.87 ± 0.00	0.75 ± 0.02
26	0.97 ± 0.04	0.93 ± 0.03	0.90 ± 0.03	0.96 ± 0.04
27	0.77 ± 0.04	0.83 ± 0.02	0.88 ± 0.01	0.78 ± 0.04
28	0.93 ± 0.02	0.95 ± 0.01	0.97 ± 0.01	0.93 ± 0.02
29	0.75 ± 0.02	0.82 ± 0.01	0.88 ± 0.00	0.76 ± 0.02
31	0.85 ± 0.02	0.91 ± 0.02	0.95 ± 0.02	0.86 ± 0.02
33	0.82 ± 0.02	0.86 ± 0.01	0.89 ± 0.00	0.83 ± 0.02
34	0.84 ± 0.04	0.87 ± 0.02	0.89 ± 0.01	0.84 ± 0.04
35	0.83 ± 0.03	0.86 ± 0.01	0.89 ± 0.00	0.83 ± 0.02
36	0.74 ± 0.02	0.85 ± 0.01	0.95 ± 0.02	0.76 ± 0.02
37	0.84 ± 0.02	0.89 ± 0.01	0.93 ± 0.02	0.85 ± 0.02
38	0.78 ± 0.04	0.84 ± 0.02	0.88 ± 0.01	0.79 ± 0.04
39	0.72 ± 0.02	0.83 ± 0.01	0.92 ± 0.02	0.74 ± 0.02
40	0.79 ± 0.03	0.84 ± 0.02	0.88 ± 0.01	0.80 ± 0.03
41	0.81 ± 0.03	0.89 ± 0.02	0.96 ± 0.02	0.82 ± 0.02
42	0.91 ± 0.03	0.84 ± 0.03	0.79 ± 0.04	0.90 ± 0.03
43	0.79 ± 0.03	0.84 ± 0.01	0.88 ± 0.00	0.80 ± 0.03
44	0.77 ± 0.01	0.87 ± 0.01	0.95 ± 0.01	0.79 ± 0.01
45	0.72 ± 0.02	0.71 ± 0.02	0.71 ± 0.02	0.72 ± 0.02
46	0.54 ± 0.02	0.57 ± 0.01	0.58 ± 0.01	0.55 ± 0.02
47	0.87 ± 0.03	0.89 ± 0.03	0.90 ± 0.03	0.87 ± 0.03
48	0.85 ± 0.04	0.87 ± 0.02	0.89 ± 0.01	0.85 ± 0.03
49	0.82 ± 0.03	0.91 ± 0.02	0.97 ± 0.02	0.83 ± 0.02
50	0.77 ± 0.05	0.83 ± 0.03	0.88 ± 0.01	0.78 ± 0.05

Continued on next page

Plane	S^2_{NH}	S^2_{CN}	S^2_{CC}	S^2_{CH}
51	0.86 ± 0.02	0.92 ± 0.02	0.96 ± 0.02	0.87 ± 0.02
52	0.89 ± 0.04	0.88 ± 0.02	0.87 ± 0.01	0.88 ± 0.03
53	0.66 ± 0.02	0.77 ± 0.01	0.86 ± 0.00	0.67 ± 0.02
54	0.70 ± 0.03	0.79 ± 0.02	0.87 ± 0.01	0.71 ± 0.03
55	0.80 ± 0.02	0.88 ± 0.02	0.95 ± 0.03	0.81 ± 0.02
56	0.95 ± 0.04	0.87 ± 0.02	0.82 ± 0.03	0.94 ± 0.04
57	0.91 ± 0.01	0.90 ± 0.01	0.90 ± 0.00	0.91 ± 0.01
58	0.86 ± 0.07	0.88 ± 0.04	0.89 ± 0.01	0.86 ± 0.06
59	0.91 ± 0.04	0.94 ± 0.02	0.96 ± 0.02	0.91 ± 0.04
60	0.91 ± 0.06	0.90 ± 0.03	0.90 ± 0.01	0.91 ± 0.06
61	0.74 ± 0.04	0.85 ± 0.02	0.95 ± 0.01	0.76 ± 0.04

Table 20 – GB3 order parameters from SF-GAF analysis. Order parameters for N_i-H^N_i, C'_{i-1}-N_i, C^α_{i-1}-C'_{i-1} and C'_{i-1}-H^N_i vector for each studied peptide plane.

Plane	σ_α	σ_β	σ_γ
8	4.60 ± 3.82	10.00 ± 1.74	14.46 ± 2.44
9	4.60 ± 4.05	10.00 ± 2.01	11.06 ± 5.55
10	4.60 ± 4.66	10.00 ± 1.86	0.00 ± 3.45
11	4.60 ± 4.32	10.00 ± 2.79	12.06 ± 3.17
12	4.60 ± 4.96	10.00 ± 2.37	16.92 ± 1.85
13	4.60 ± 4.62	10.00 ± 2.67	0.01 ± 4.93
14	4.60 ± 4.54	14.18 ± 0.82	0.00 ± 2.60
15	4.60 ± 3.76	12.32 ± 1.05	18.24 ± 2.75
16	4.60 ± 4.96	15.83 ± 1.85	14.31 ± 1.65
17	4.60 ± 3.90	10.00 ± 2.14	29.25 ± 1.91
18	4.60 ± 3.90	0.00 ± 3.46	18.22 ± 1.10
19	4.60 ± 4.03	10.00 ± 2.32	22.84 ± 3.04
20	4.60 ± 3.98	10.00 ± 2.43	16.43 ± 4.45
21	17.18 ± 2.23	9.61 ± 1.87	11.14 ± 3.73
22	4.60 ± 4.23	10.00 ± 2.67	6.78 ± 6.81
23	4.60 ± 3.93	10.00 ± 2.69	11.87 ± 1.32
24	4.60 ± 4.02	10.00 ± 1.59	17.37 ± 1.77
25	4.60 ± 4.48	10.00 ± 4.45	15.76 ± 1.17

Continued on next page

Plane	σ_α	σ_β	σ_γ
26	9.86 ± 2.41	5.27 ± 2.78	0.00 ± 3.78
27	4.60 ± 4.17	10.00 ± 1.94	13.86 ± 2.44
28	4.60 ± 3.78	0.01 ± 2.43	9.14 ± 3.97
29	4.60 ± 6.45	10.00 ± 3.51	15.20 ± 1.25
31	4.60 ± 4.13	4.35 ± 3.33	13.09 ± 0.95
33	4.60 ± 4.18	10.00 ± 2.50	10.89 ± 1.09
34	4.60 ± 4.61	10.00 ± 2.71	9.78 ± 3.80
35	4.60 ± 4.26	10.00 ± 3.91	10.46 ± 1.93
36	4.60 ± 5.20	0.00 ± 3.06	19.22 ± 1.56
37	4.60 ± 5.54	6.76 ± 3.04	12.43 ± 1.68
38	4.60 ± 5.65	10.00 ± 3.82	13.36 ± 3.60
39	4.60 ± 3.78	5.44 ± 2.95	18.94 ± 1.52
40	4.60 ± 5.80	10.00 ± 2.71	12.96 ± 2.00
41	4.60 ± 4.65	0.00 ± 2.83	16.11 ± 1.57
42	13.31 ± 1.92	10.04 ± 4.38	0.00 ± 5.33
43	4.60 ± 3.76	10.00 ± 2.34	12.77 ± 1.66
44	4.60 ± 4.50	0.00 ± 2.78	17.67 ± 1.11
45	4.60 ± 7.10	19.44 ± 0.93	0.00 ± 3.40
46	4.60 ± 5.91	24.77 ± 0.78	10.24 ± 2.05
47	10.12 ± 2.07	2.90 ± 3.23	12.42 ± 1.21
48	4.60 ± 4.18	10.00 ± 1.59	8.94 ± 3.59
49	0.00 ± 2.58	0.00 ± 2.85	15.49 ± 1.04
50	4.60 ± 04.28	10.00 ± 2.05	14.09 ± 4.25
51	4.60 ± 3.56	0.00 ± 3.20	13.34 ± 0.89
52	4.60 ± 4.32	11.59 ± 1.23	0.00 ± 5.25
53	4.60 ± 5.75	10.00 ± 2.23	19.87 ± 1.04
54	4.60 ± 5.28	10.00 ± 3.16	17.82 ± 1.62
55	4.60 ± 4.14	2.63 ± 3.57	16.24 ± 0.99
56	13.66 ± 1.91	7.04 ± 2.66	0.00 ± 5.23
57	4.60 ± 4.07	10.00 ± 2.57	2.18 ± 2.66
58	4.60 ± 3.95	10.00 ± 2.37	8.26 ± 5.53
59	4.60 ± 4.57	4.21 ± 2.92	9.75 ± 3.35
60	4.60 ± 3.84	10.00 ± 2.48	0.00 ± 5.95
61	4.60 ± 4.64	0.00 ± 2.75	19.02 ± 2.12

Table 21 – Order parameters obtained through `3-1D-GAF` analysis for GB3 β-sheet.

	Shared motion				Local motion			
Plane	S^2_{NH}	S^2_{CN}	S^2_{CC}	S^2_{CH}	S^2_{NH}	S^2_{CN}	S^2_{CC}	S^2_{CH}
9	0.90	0.90	0.90	0.90	0.97	0.98	1.00	0.97
10	0.89	0.92	0.90	0.90	1.00	1.00	1.00	1.00
11	0.89	0.90	0.90	0.89	0.88	0.94	0.98	0.89
12	0.89	0.91	0.92	0.89	0.85	0.93	0.98	0.86
13	0.89	0.89	0.90	0.89	1.00	1.00	1.00	1.00
18	0.89	0.91	0.93	0.89	0.85	0.92	0.98	0.86
19	0.89	0.91	0.89	0.90	0.71	0.84	0.96	0.73
20	0.89	0.89	0.90	0.89	0.84	0.92	0.98	0.85
21	0.89	0.90	0.90	0.89	0.92	0.96	0.99	0.93
22	0.91	0.89	0.91	0.90	0.93	0.96	0.99	0.93
23	0.93	0.90	0.92	0.91	0.92	0.96	0.99	0.93
47	0.90	0.94	0.91	0.93	0.94	0.97	0.99	0.94
48	0.90	0.90	0.91	0.90	0.93	0.97	0.99	0.94
49	0.89	0.92	0.93	0.90	0.90	0.95	0.99	0.91
50	0.89	0.92	0.90	0.91	0.91	0.95	0.99	0.91
51	0.90	0.91	0.94	0.89	0.93	0.96	0.99	0.94
56	0.91	0.89	0.91	0.90	1.00	1.00	1.00	1.00
57	0.89	0.94	0.91	0.92	1.00	1.00	1.00	1.00
58	0.89	0.89	0.90	0.89	0.94	0.97	0.99	0.94
59	0.89	0.91	0.93	0.90	1.00	1.00	1.00	1.00
60	0.89	0.89	0.90	0.89	1.00	1.00	1.00	1.00

D

WEAK COMPLEX FORMATION IN THE HIGHLY DILUTED LIMIT

The aim of the annexe is to investigate, for the case of a weak complex, the possibility to saturate one of the two partners without using excessively high concentrations of the other.

Considering a complex formation, defined by the following equilibrium and concentrations:

	$\mathbb{A}$	$+$	$\mathbb{B}$	$\rightleftharpoons$	$\mathbb{C}$
Introduced Concentration	$[\mathbb{A}]_0$		$[\mathbb{B}]_0$		0
Effective Concentration	$[\mathbb{A}] = [\mathbb{A}]_0 - [\mathbb{C}]$		$[\mathbb{B}] = [\mathbb{B}]_0 - [\mathbb{C}]$		$[\mathbb{C}]$

where the associated dissociation constant $\mathbb{K}_d$, is given by:

$$\mathbb{K}_d = \frac{[\mathbb{A}][\mathbb{B}]}{[\mathbb{C}]C^0} \qquad \text{or} \qquad K_d = \frac{[\mathbb{A}][\mathbb{B}]}{[\mathbb{C}]} \qquad \text{with} \qquad C^0 = 1\,\text{mol}\cdot\text{L}^{-1} \tag{D.1}$$

which becomes, using the initial concentrations:

$$K_d = \frac{([\mathbb{A}]_0 - [\mathbb{C}])([\mathbb{B}]_0 - [\mathbb{C}])}{[\mathbb{C}]} \tag{D.2}$$

Thus, the concentration of the complex can be found by solving the equation:

$$[\mathbb{C}]^2 - (K_d + [\mathbb{A}]_0 + [\mathbb{B}]_0)[\mathbb{C}] + [\mathbb{A}]_0[\mathbb{B}]_0 = 0 \tag{D.3}$$

Roots of this second-order equation can be found as:

$$[\mathbb{C}] = \frac{K_d + [\mathbb{A}]_0 + [\mathbb{B}]_0 \pm \sqrt{\Delta}}{2} \qquad \text{with} \qquad \Delta = (K_d + [\mathbb{A}]_0 + [\mathbb{B}]_0)^2 - 4[\mathbb{A}]_0[\mathbb{B}]_0 \tag{D.4}$$

The only physically acceptable solution is the one that, in absence of one of the two partners $\mathbb{A}$ or $\mathbb{B}$ leads to the nonexistence of the complex, which is:

$$[\mathbb{C}] = \frac{K_d + [\mathbb{A}]_0 + [\mathbb{B}]_0 - \sqrt{(K_d + [\mathbb{A}]_0 + [\mathbb{B}]_0)^2 - 4[\mathbb{A}]_0[\mathbb{B}]_0}}{2} \tag{D.5}$$

Now, the concentration of the partner $\mathbb{A}$ ($[\mathbb{A}]_0$) is fixed and the effect of diluting $\mathbb{B}$ ($[\mathbb{B}]_0$) on the complex formation will be investigated. The problem is completely symmetric as $\mathbb{A}$ and $\mathbb{B}$ can be exchanged. We have:

$$[\mathbb{C}] = \frac{K_d + [\mathbb{A}]_0 + [\mathbb{B}]_0}{2}\left(1 - \sqrt{1 - \frac{4[\mathbb{A}]_0[\mathbb{B}]_0}{(K_d + [\mathbb{A}]_0 + [\mathbb{B}]_0)^2}}\right) \tag{D.6}$$

In order to characterize the fraction of $\mathbb{B}$ in the complex, we introduce:

$$P_{\mathbb{B}} = \frac{[\mathbb{C}]}{[\mathbb{B}]_0} \tag{D.7}$$

which for $[\mathbb{B}]_0$ tending towards zero is to a first order approximation given by:

$$P_{\mathbb{B}} = \frac{K_d + [\mathbb{A}]_0 + [\mathbb{B}]_0}{2[\mathbb{B}]_0}\left(1 - \left(1 - \frac{2[\mathbb{A}]_0[\mathbb{B}]_0}{(K_d + [\mathbb{A}]_0 + [\mathbb{B}]_0)^2}\right)\right) \tag{D.8}$$

which leads to:

$$P_{\mathbb{B}} = \frac{[\mathbb{A}]_0}{K_d + [\mathbb{A}]_0 + [\mathbb{B}]_0} \tag{D.9}$$

As we are looking for the limiting value of $P_{\mathbb{B}}$ when $[\mathbb{B}]_0$ tends towards zero, we have:

$$\lim_{[\mathbb{B}]_0 \to 0} P_{\mathbb{B}} = \frac{[\mathbb{A}]_0}{K_d + [\mathbb{A}]_0} \tag{D.10}$$

which demonstrates that by diluting $\mathbb{B}$ it will be impossible to reach $P_{\mathbb{B}} = 1$ which corresponds to the saturation of $\mathbb{B}$, if the concentration of $\mathbb{A}$ is not much bigger than the K_d constant.

Considering a weak interacting complex, e.g. $K_d = 0.2\,\mathrm{mM}$, and we want to limit the concentration of the partners at reasonably low concentrations (for biological relevance or to avoid oligomerization or aggregation), e.g. $[\mathbb{A}]_0 = 1\,\mathrm{mM}$, the $\mathbb{B}$ compound can never be saturated, the maximal fraction in the complex being: $P_{\mathbb{B}}$ =0.83. Using a high concentration of $[\mathbb{A}]_0 = 5\,\mathrm{mM}$ will only lead to a maximal $P_{\mathbb{B}}$ =0.96. Therefore weak complexes cannot be properly studied by direct saturation and titration approaches are required.

E

RÉSUMÉ EN FRANÇAIS

E.1 INTRODUCTION

La compréhension des organismes vivants a occupé une place importante dans le développement des concepts scientifiques et philosophiques du fait, certes, de son incroyable complexité mais surtout car elle rejoint les interrogations et les réflexions visant à définir l'être humain et à comprendre son existence.

L'approche épistémologique proposée ici peut être perçue comme une approche réductionniste, dans le sens ouvert et positif du terme, proposant d'utiliser au mieux les connaissances issues de la Physique et de la Chimie afin d'apporter un éclairage nouveau sur les Sciences Biologiques.

Le présent travail s'intéresse à l'étude du désordre conformationnel dans les protéines, molécules essentielles au fonctionnement du Vivant. L'importance de ce désordre conformationnel commence à être perçu comme un moyen d'outrepasser les limites d'une description complètement structurale de la Biologie qui a permis au cours des dernières décennies des avancées extrêmement importantes dans la compréhension des systèmes vivants.

La Résonance Magnétique Nucléaire est utilisée ici comme outil principal pour étudier cette dynamique. Deux raisons motivent ce choix : tout d'abord la possibilité d'obtenir de l'information spécifique résolue à l'échelle atomique et ensuite la sensibilité des mesures RMN à de nombreuses échelles de temps. L'interaction la plus utilisée dans ces travaux est le Couplage Dipolaire Résiduel (RDC) qui permet d'obtenir de l'information à des échelles de temps allant jusqu'à la milliseconde.

Cette Thèse est organisée en trois parties : une partie théorique pose les bases de la physique nécessaires au développement des méthodes décrites par la suite, une seconde partie se concentre sur la détermination quantitative de la dynamique des protéines repliées et enfin la dynamique conformationnelle des protéines intrinsèquement désordonnées est décrite dans une dernière partie.

E.2 CONCEPTS THÉORIQUES

E.2.1 *La Relaxation en RMN*

La relaxation de spin consiste un outil extrêmement utile pour sonder la dynamique rapide des protéines. En effet, elle caractérise les processus induisant un retour à l'équilibre de l'aimantation mise hors équilibre au cours d'une expérience RMN. Les mécanismes de relaxation sont induits par des mouvements stochastiques produisant des modulations dans la géométrie du système de spin, ces fluctuations induisant les transitions nécessaires au retour à l'équilibre.

Le formalisme de l'opérateur densité est introduit, permettant de décrire par cet unique opérateur l'évolution des cohérences et des populations présentes pour chaque état ou entre les différents états de spins. Basée sur cet outil et sur une description semi-classique du système étudié — le système de spin étant traité de manière quantique et le réseau de façon classique — la théorie de Redfield-Abragam permet de dériver l'équation maîtresse qui caractérise l'évolution du système de spin. Cette approche introduit deux fonctions extrêmement importantes : la fonction d'auto-corrélation et la fonction de densité spectrale, ces deux fonctions, l'une de variable temporelle, l'autre de fréquence, étant liées par transformée de Fourier. La fonction d'auto-corrélation caractérise la perte de mémoire pour un système évoluant aléatoirement, la fonction de densité spectrale représentant la distribution de fréquences présente dans le système considéré.

Les mécanismes de relaxation ^{15}N, d'origine dipolaire ou basée sur l'interaction d'anisotropie de déplacement chimique, et l'effet de l'échange chimique sont exposés.

Enfin l'extraction d'information sur le mouvement étudié est présentée principalement dans le cadre de l'analyse dite model-free dans laquelle aucune forme spécifique du mouvement n'est postulée. Ce dernier est décrit par deux grandeurs, un paramètre d'ordre qui caractérise l'extension spatiale du mouvement et un temps de corrélation qui décrit la vitesse de décorrélation associée au mouvement étudié.

E.2.2 *Les Couplages Dipolaires Résiduels*

Les couplages dipolaires résiduels dérivent de l'interaction dipolaire existante entre deux moments magnétiques. Le mouvement moléculaire moyenne l'interaction dipolaire et l'utilisation de milieux orientants permet

d'obtenir l'information contenue dans cette interaction tout en conservant la simplicité des spectres observables en RMN à l'état liquide.

Les aspects expérimentaux de l'utilisation de milieux orientants sont présentés. Les manifestations de l'anisotropie de déplacement chimique, des couplages dipolaires et quaduroplaires sur les spectres RMN sont décrites et les différents types de milieux orientants existants — alignement magnétique, cristaux liquides ou gels déformés — sont présentés.

Les expressions analytiques des RDCs sont ensuite dérivées du couplage dipolaire et discutées dans le cadre de l'approximation séculaire ou de la limite des couplages faibles. Les hypothèses d'un découplage du mouvement moléculaire global et de la dynamique interne permettent d'introduire le formalisme de la matrice de Saupe et d'ensuite décrire la dynamique locale dans un repère moléculaire lié à la protéine considérée. Cette dynamique locale peut être décrite à l'aide des harmoniques sphériques et du formalisme des rotations d'Euler permettant d'exprimer ce moyennement dynamique dans n'importe quel système de coordonnées locales et d'introduire un paramètre d'ordre, comparable à celui issu de la relaxation de spin caractérisant l'amplitude de la dynamique présente dans le système étudié.

Dans le cadre d'une interprétation purement statique des RDCs, une importante quantité d'information structurale peut être obtenue, malgré une intrinsèque dégénérescence de l'information obtenue. Cette information est caractérisée pour différents types de systèmes, en particulier ceux importants pour la description des systèmes protéiques : les vecteurs isolés, les systèmes plans et les objets chiraux.

Enfin l'effet du moyennement dynamique des RDCs est détaillé pour différents types de mouvements comme la réorientation dans un cône ou le mouvement des Fluctuations Axiales Gaussiennes (GAF).

E.3 DYNAMIQUE DES PROTÉINES REPLIÉES

E.3.1 *Tour d'Horizon des Descriptions Dynamiques des RDCs*

Les RDCs constituant d'importantes sondes de dynamique, la mise en place de formalismes permettant de décoder des mesures expérimentales les caractéristiques du mouvement sont nécessaires. Dans un premier temps, les limites d'un modèle purement structural sont exposées et une méthode mathématique quantifiant l'auto-cohérence d'un jeu expérimental de RDC est décrite.

Ensuite les méthodes existantes permettant de décrire la dynamique contenue dans les RDC sont présentées. Tout d'abord les approches basées sur les analyses de tenseurs de fragments sont exposées, soit pour des domaines entiers, soit pour de plus petits fragments, tel qu'un plan peptidique et de la jonction tetrahèdrique suivante, comme dans l'analyse de tenseur d'alignement local.

Les méthodes de moyennes d'ensemble sont ensuite introduites, aussi bien pour les dynamiques moléculaires sous contraintes combinant en un seul potentiel les données expérimentales et un champ de force ou les approches basées purement sur la dynamique moléculaire, dont les trajectoires fournissent une description intuitive du mouvement. Enfin les approches combinant un premier échantillonnage large suivi d'une sélection d'un sous-ensemble sur la base des données expérimentales sont décrites.

Par ailleurs, les approches de type model-free sont introduites via les approches SCRM (Self-Consistant RDC-based Model-Free Approoach) et DIDC (Direct Analysis of Dipolar Couplings). Ces approches permettent une description détaillée du mouvement moléculaire en caractérisant, sans hypothèse de mouvement, le moyennement des harmoniques sphériques présentes dans l'expression analytique des RDCs moyennés dynamiquement.

Enfin l'ensemble des travaux précédents cette Thèse sur les modèles GAF sont présentés.

E.3.2 *Détermination Quantitative et Absolue de la Dynamique de la Chaîne Principale de l'Ubiquitine*

La détermination du niveau absolu de dynamique dans les protéines reste une question ouverte, cruciale pour la compréhension de l'importance de ces mouvements dans les systèmes biologiques. Jusqu'à présent la plupart des méthodes analytiques utilisaient la relaxation ^{15}N de spin pour fixer la gamme de dynamique. Ici une méthode est développée pour déterminer uniquement à partir des RDCs, la quantité de dynamique présente dans l'Ubiquitine. Cette approche est basée sur le modèle GAF, ici exprimé dans un formalisme efficace et complètement dédié à la dynamique. Cette approche permet d'obtenir via des analyses indirectes un niveau de dynamique, qui par comparaison à des analyses menées sur des données simulées, peut être considéré comme quantitatif.

Un modèle complet de mouvement visant à éviter une surinterprétation des données a été développé. Les informations apportées par cette analyse ont

été comparées aux analyses de relaxation ^{15}N, à des dynamiques moléculaires et à l'approche SCRM. Ces résultats sont cohérents avec l'absence de mouvements importants dans les structures secondaires et à l'inverse une claire présence de mouvements lents dans les boucles et les parties les moins structurées est visible. De plus un mouvement lent présent dans une boucle en C-terminal, importante pour la reconnaissance moléculaire, présente un mouvement lent, détecté seulement par les approches utilisant les RDCs.

L'information structurale obtenue au cours de cette analyse permet de caractériser la position moyenne de chaque vecteur d'un plan peptidique et l'incertitude associée. Le modèle a été testé, passant successivement de nombreux tests statistiques, aussi bien pour des analyses directes que pour des validations croisées. Tous ces tests convergent vers une très haute pertinence de ce modèle, estimée sur la base d'un test AIC comme plus d'un million de fois plus probable qu'une analyse statique.

E.3.3 *Dynamique Moléculaire Accélérée de l'Ubiquitine*

La possibilité d'obtenir des ensembles moléculaires afin de décrire la dynamique des protéines à des échelles de temps lents donnerait une vue complémentaire de celle proposée par la description analytique donnée par les approches de type GAF. Néanmoins la puissance de calcul actuellement disponible n'est pas en mesure de générer des trajectoires qui échantillonnent de manière significative des mouvements aux échelles de temps allant jusqu'à la milliseconde. L'approche proposée ici, basée sur les méthodes de dynamique moléculaire accélérée permet d'échantillonner ses mouvements en biaisant le potentiel auquel le système est soumis.

L'approche consiste à tout d'abord biaiser le potentiel, par une transformation à deux variables, permettant de conserver la rugosité du potentiel tout en faisant décroître l'amplitude des barrières d'énergie potentielle. L'ensemble ainsi obtenu sert de point de départ à des dynamiques moléculaires standard, permettant d'obtenir des ensembles pesés selon une distribution boltzmannienne. Ce processus — accélération puis dynamique moléculaire standard — est répété pour différents niveaux d'accélération permettant d'obtenir différents ensembles correspondant à des échelles de temps plus ou moins longues. La sélection du niveau d'accélération se fait simplement en sélectionnant le niveau d'accélération conduisant à la meilleure reproduction des données expérimentales considérées, celle-ci devant être sensible à l'échelle de temps étudiée. Ici, l'étude se focalisait sur les échelles de temps allant jusqu'à la milliseconde et c'est pourquoi les données expérimentales utilisées ont été des RDCs et des couplages scalaires, tous deux sensibles à cette gamme de temps.

L'ensemble moléculaire ainsi obtenu a été comparé à diverses approches, principalement à la méthode SF-GAF. La convergence des paramètres d'ordre obtenus par les deux approches est globalement très bonne — relativement bonne pour les paramètres d'ordre C^{α}_{i-1}-C'_{i-1} et excellente pour les paramètres d'ordre N_i-H^N_i—. Cette très bonne convergence, présente malgré des approches portées par hypothèses très différentes, renforce certainement la confiance que l'on peut placer en ces deux descriptions.

E.3.4 *Étude de la Dynamique de la Protéine GB3 : vers une Description des Mouvements Collectifs*

Les descriptions analytiques permettant d'extraire de l'information dynamique des RDCs se basent sur des descriptions purement locales du mouvement. Par exemple la description GAF du mouvement est locale — spécifique à chaque plan peptidique — et traite le mouvement de manière complètement décorrélé et individuel. Néanmoins les mouvements collectifs et — ou — corrélés peuvent avoir un rôle majeur au niveau biologique et être partiellement voire complètement masqué par ce type de description.

Dans ce chapitre, différents modèles généralisant la description GAF, sont développés afin de prendre en compte explicitement un mouvement collectif — ou commun — et un mouvement local. Pour le système étudié, la protéine GB3, la possibilité d'un mouvement corrélé au sein du feuillet β a en effet été mise en évidence au cours d'étude par RMN en phases liquide et solide.

Une première analyse de type SF-GAF a été réalisée pour ce système, permettant de déterminer le niveau de dynamique présent dans cette protéine. Cette analyse présente des résultats très similaires à celle menée sur l'Ubiquitine, indiquant une éventuelle généralité des résultats obtenus dans la première analyse. Après les développements analytiques nécessaires à la mise en place de ces modèles, ils ont pu être appliqués au feuillet β et à l'hélice α de la protéine GB3. Dans les deux cas un mouvement anisotrope collectif a pu être mis en évidence, mais il n'a pu être validé que dans le cas du feuillet β. Pour ce motif une description intégrant explicitement les mouvements locaux et communs a été utilisé. L'utilisation d'une description de la dynamique locale par un modèle de type 3D-GAF ne laisse pas apparaître de mouvement commun même si la présence d'un tel mouvement de faible amplitude ne peut être exclu. En revanche, si le mouvement local est limité à une réorientation de type γ-1D-GAF, un mouvement collectif anisotrope peut être détecté tant par des analyses directes qu'indirectes. La combinaison de ces deux mouvements permet d'obtenir des distributions de paramètres d'ordre extrêmement similaire à celles obtenues par une analyse complètement locale de type SF-GAF.

Ces approches, prenant explicitement en compte ces deux types de dynamique, semblent pouvoir s'appliquer avec pertinence à des systèmes dynamiques plus complexes, tels que les complexes protéines-protéines ou les acides nucléiques.

E.3.5 *Étude de la Dynamique Rapide et Lente de la Protéine SH3-C*

Les précédents chapitres se sont concentrés sur la dynamique de deux systèmes : les protéines Ubiquitine et GB3. Néanmoins les résultats sont encore trop peu nombreux pour pouvoir généraliser ses approches et l'étude de nouveaux systèmes, avec des propriétés différentes apporterait une information cruciale sur la généralité des résultats précédemment obtenus. Le domaine SH3-C étudié ici, présente un repliement complètement différent des systèmes jusqu'alors étudiés, et permet donc d'aborder un système différent.

Pour ce système des données de relaxation ont été enregistrées et ont permis une analyse de sa dynamique rapide, incluant la détermination du tenseur de diffusion et de la mobilité interne. De plus, des RDCs ont été mesurés dans 15 milieux orientants différents, avec jusqu'à trois couplages par plan peptidique. De ce vaste jeu de données, un sous-ensemble de 10 milieux a été sélectionné pour son auto-cohérence par une approche de type SECONDA. Ce jeu de RDC a permis de mener une analyse DYNAMIC-MECCANO suivi d'un `3D-GAF` sur cette structure, permettant une détermination simultanée de la structure et de la dynamique du domaine considéré. La structure DYNAMIC-MECCANO présente des caractéristiques très proches de celles obtenues par une description purement statique de cet ensemble de RDC. La dynamique observable dans ce système peut d'ailleurs être corrélée avec la présence de variabilité conformationnelle dans différents domaines SH3 dont les structures, en complexe avec différents partenaires, ont été déterminées par rayon X.

Afin d'étudier plus en détail cette variabilité conformationnelle une approche permettant de sélectionner un ensemble moléculaire a été mise au point. Basée sur un algorithme génétique, nommé ASTEROIDS et initialement développé pour les protéines intrinsèquement désordonnées, cette méthode permet de sélectionner dans une large base de donnée — ici constituée des structures issues d'une longue trajectoire de dynamique moléculaire (1 µs) de ce domaine SH3 — un sous-ensemble, représentatif des données expérimentales considérées. Ici le jeu de RDCs précédemment obtenu a été utilisé et une méthode utilisant la décomposition en valeur singulière a été implémentée pour confronter l'échantillonnage conformationnel du sous-ensemble considéré aux RDCs mesurés.

Même si des études sont toujours en cours sur ces systèmes, les résultats déjà obtenus semblent converger vers une situation assez similaire à celles des protéines Ubiquitine et GB3 où peu de dynamique lente était détectable dans les structures secondaires mais où les boucles présentaient des mouvements à des échelles de temps plus lentes. De plus une importante dynamique conformationnelle a été révélée pour la protéine seule, en solution, dans les régions en interactions avec des partenaires biologiques.

E.3.6 *Étude du Complexe Faible SH3-C Ubiquitine par Relaxation ^{15}N*

L'étude des complexes faibles reste un des problèmes difficilement abordables par les différentes méthodes biophysiques car il est souvent impossible d'isoler l'entité constituée des deux partenaires en interaction. L'utilisation de méthode de titration où l'information mesurée résulte à la fois de la forme libre et de la forme en complexe, peut permettre par des processus d'extrapolation d'isoler des grandeurs caractéristiques uniquement du complexe.

L'étude menée ici a tout d'abord permis de déterminer la constante thermodynamique associée à la dissociation du complexe. La faiblesse de cette interaction ne permettant pas d'isoler le complexe dans des conditions raisonnables, les vitesses de relaxation longitudinale (R_1) ou transverse (R_2) ont du elles aussi être extrapolées dans le complexe. Considérant un modèle cinétique à deux sites, l'évolution des R_1 est supposée être linéaire. Un comportement similaire est attendu pour les R_2 où aucun échange n'est présent, par contre une contribution supplémentaire est attendue pour les sites ne présentant pas le même environnement chimique dans la forme libre et dans le complexe. Cette contribution d'échange a été estimée en utilisant une analyse de type model-free pour chaque protéine dans chaque mélange. En retirant cette contribution de l'évolution des R_2 expérimentaux, les R_2 intrinsèques sont obtenus et peuvent être ainsi extrapolés dans le complexe.

Les vitesses de relaxation obtenues dans le complexe ont servi à déterminer le tenseur d'alignement de cette espèce, conduisant à un tenseur de symétrie axial, hautement anisotrope. La dépendance angulaire des rapports R_2/R_1 a été comparée à celle de RDCs, mesurés dans un milieu orientant stérique, et extrapolés dans le complexe. La corrélation présente entre ses variations renforce la validité des deux procédures. Enfin la dynamique rapide des deux protéines dans le complexe a pu être déterminée et comparée à celles des formes libres. Peu de variations ont pu être observées en termes de paramètre d'ordre, à l'exception d'une rigidification présente légèrement dans certaines zones d'interaction et plus clairement dans la partie C-

terminal de l'Ubiquitine, connue pour jouer un rôle clef dans la formation du complexe.

E.4 PROTÉINES INTRINSÈQUEMENT DÉSORDONNÉES ET ASTEROIDS

E.4.1 *Les Protéines Chimiquement Dénaturées et Intrinsèquement Désordonnées comme Systèmes Extrêmement Flexibles*

Les protéines intrinsèquement désordonnées (IDPs) sont maintenant reconnues comme des systèmes jouant des rôles clefs dans le fonctionnement du vivant. Du fait de leur plasticité intrinsèque, ces protéines peuvent interagir avec différents partenaires, se replier aux cours d'interaction avec un partenaire ou présenter de très importantes surfaces de contact sans nécessité des poids moléculaires extrêmes. Ces protéines qui peuvent représenter jusqu'à 40 % du protéome humain sont impliquées par exemple dans la régulation du cycle cellulaire, dans la signalisation ou dans des maladies telles que les maladies neurodégénératives et le cancer.

La RMN constitue une méthode de choix pour étudier ces systèmes car elle permet d'accéder à de l'information aussi bien locale qu'à longue portée sur ces systèmes extrêmement flexibles. Même si par exemple les déplacements chimiques ou les mesures de relaxation ont pu apporter des informations intéressantes sur ces systèmes, les deux sources d'information les plus utilisées sont pour l'instant les RDCs et les PREs (Paramagnetic Relaxation Enhancement).

L'approche FLEXIBLE-MECCANO permet de décrire des protéines dans un état complètement désordonné en supposant un échantillonnage conformationnel local suivant les distributions dites random-coil, obtenu en ne retenant d'une large base de données issues de 500 structures haute résolution que les conformations ne faisant partie ni d'hélice α ni de feuillet β. Cette description est dépendante de la séquence primaire et permet de donner une bonne représentation de la dynamique conformationnelle de systèmes complètement désordonnés. En utilisant simplement cette description, tout ordre résiduel à longue ou à courte portée doit être caractérisé par des méthodes biaisant l'échantillonnage de départ.

E.4.2 *Caractérisation de l'Ordre Local dans les Systèmes Désordonnés*

L'existence d'ordre local dans les protéines désordonnées a été mise en évidence expérimentalement par diverses approches. L'idée de ce chapitre est

de développer une approche permettant de caractériser cet ordre résiduel sans introduire d'hypothèse pouvant biaiser l'analyse du système. À partir d'un échantillonnage obtenu par FLEXIBLE-MECCANO, un algorithme génétique, appelé ASTEROIDS a été développé pour sélectionner un sous-ensemble de structures permettant de reproduire les données expérimentales et les propriétés biophysiques des systèmes considérés. Si la première partie est relativement facile à réaliser la seconde s'avère beaucoup plus difficile car le problème à résoudre est sous déterminé, en ce sens que les données expérimentales disponibles ne permettent pas de définir complètement l'échantillonnage conformationnel de chaque membre du sous-ensemble étudié. De ce fait, une importante série de tests a été mise en place afin de déterminer les conditions dans lesquelles la reproduction des données expérimentales implique une caractérisation satisfaisante des propriétés biophysiques du système considéré.

Cette approche a d'abord été développée pour étudier l'Ubiquitine dénaturée dans l'urée à l'aide de RDCs. Cette étude a permis de déterminer l'échantillonnage conformationnel de chaque acide aminé et d'en caractériser la déviation par rapport à l'état de référence qu'est le random-coil. Ensuite cette méthode a été utilisée pour caractériser N_{TAIL} à l'aide des déplacements chimiques uniquement, permettant de mettre en évidence l'existence de structures secondaires résiduelles dans la forme libre de cette entité, connue pour subir une transition désordre-ordre lors de l'interaction avec son partenaire biologique.

E.4.3 *Caractérisation de l'Ordre à Longue Portée dans les Systèmes Désordonnés*

L'existence d'ordre à longue portée dans les systèmes intrinsèquement désordonnés peut se révéler d'une grande importance aussi bien pour le rôle biologique de telles protéines que pour l'interprétation physique de grandeurs telles que les RDCs.

Ici l'algorithme ASTEROIDS a été utilisé pour étudier l'ordre à longue portée dans l'α-Synuclein en utilisant des PREs. En effet, l'amplification des vitesses de relaxation liée à la présence d'un noyau paramagnétique est directement dépendante de la distribution de distance entre ce noyau paramagnétique et le spin étudié et constitue donc une intéressante sonde des interactions transitoires présentes dans ses systèmes dynamiques.

Cette approche a là encore été testée avec des données simulées avant être appliquée à l'α-Synuclein, permettant de révéler l'existence d'interactions à longue portée entre l'extrémité N- et C-terminal et dans une moindre

mesure entre le C-terminal et le domaine NAC, présent au centre de la protéine.

L'incidence de ces interactions à longue portée sur les RDCs a été examinée, permettant de mettre en évidence de nettes distorsions du profil attendu pour les RDCs en fonction de l'existence et de la localisation de ces contacts. Cette dépendance a été paramétrée et combinée avec succès à une description locale de l'échantillonnage conformationnel, permettant de reproduire, en combinant cette paramétrisation et un nombre restreint de conformères, les RDCs expérimentaux de l'α-Synuclein mieux qu'en utilisant un très large nombre de conformères où aucun ordre à longue portée n'était supposé.

E.4.4 *Conclusion*

Cette Thèse s'est concentrée sur la caractérisation du désordre conformationnel, de manière quantitative, dans différents systèmes, aussi bien dans les protéines repliées que pour les systèmes désordonnés. De manière intéressante, la relation existant entre ces deux types de système n'est pas si opposée que ne le laisse paraître le premier abord.

En effet, l'étude de la dynamique dans les systèmes repliés peut être vue comme un écart à la situation idéale que représente une description complètement statique du système considéré. De manière symétrique, l'étude des protéines intrinsèquement désordonnées tend à caractériser ces systèmes par leur déviation par rapport à l'état complètement déplié qu'est le random-coil. Ainsi pour ses deux types de systèmes, la complémentarité de l'ordre et du désordre apparaît clairement.

Même si le gouffre entre les protéines repliées et dépliées n'est pas encore comblé et même si la caractérisation des protéines intrinsèquement désordonnées n'est pas encore assez avancée pour définir un nouveau paradigme, complet et cohérent, la complémentarité — plus que la dichotomie — entre ces deux aspects commence à clairement se révéler.

C'est pourquoi, je pense, le paradigme naissant de la biophysique du désordre dynamique (est-ce ainsi qu'il faut le nommer?), n'a pas à être construit en opposition à celui de la biologie structurale, mais devrait engendrer un cadre conceptuel plus large dans lequel l'entrelacement entre ordre et désordre conformationnel pourrait pleinement s'exprimer.

F

PUBLICATIONS

Hus JC, Salmon L, Bouvignies G, Lotze J, Blackledge M, Brüschweiler R, 16-Fold Degeneracy of Peptide Plane Orientations from Residual Dipolar Couplings: Analytical Treatment and Implications for Protein Structure Determination, *J. Am. Chem. Soc.* 2008, 130: 15927-15937

Salmon L, Bouvignies G, Markwick PRL, Lakomek N, Showalter S, Li DW, Walter K, Griesinger C, Brüschweiler R, Blackledge M, Protein Conformational Flexibility from Structure-Free Analysis of NMR Dipolar Couplings: Quantitative and Absolute Determination of Backbone Motion in Ubiquitin, *Angew. Chem. Int. Ed.* 2009, 48: 4154-4157

Markwick PRL, Bouvignies G, Salmon L, McCammon JA, Nilges M, Blackledge M, Toward a Unified Representation of Protein Structural Dynamics in Solution, *J. Am. Chem. Soc.* 2009, 131: 16968-16975

Nodet G, Salmon L, Ozenne V, Meier S, Jensen MR, Blackledge M, Quantitative Description of Backbone Conformational Sampling of Unfolded Proteins at Amino Acid Resolution from NMR Residual Dipolar Couplings, *J. Am. Chem. Soc.* 2009, 131: 17908-17918

Jensen MR, Salmon L, Nodet G, Blackledge M, Defining Conformational Ensembles of Intrinsically Disordered and Partially Folded Proteins Directly from Chemical Shifts, *J. Am. Chem. Soc.* 2010, 132: 1270-1271

Salmon L, Nodet G, Ozenne V, Yin G, Jensen MR, Zweckstetter M, Blackledge M, NMR Characterization of Long-Range Order in Intrinsically Disordered Proteins, *J. Am. Chem. Soc.* 2010, 132: 8407-8418

Salmon L, La biophysique des protéines, un domaine où règne le plus grand des ordres, Chapter from Ordre et Deśordre *L'Harmattan* 2009

J|A|C|S
ARTICLES

Published on Web 10/30/2008

16-Fold Degeneracy of Peptide Plane Orientations from Residual Dipolar Couplings: Analytical Treatment and Implications for Protein Structure Determination

Jean-Christophe Hus,[#,†,‡] Loïc Salmon,[‡] Guillaume Bouvignies,[‡] Johannes Lotze,[‡] Martin Blackledge,[*,‡] and Rafael Brüschweiler[*,§]

Clark University, Worcester, Massachusetts 01610, Institut de Biologie Structurale Jean-Pierre Ebel, 38027 Grenoble, France, and Chemical Sciences Laboratory, Department of Chemistry and Biochemistry, and National High Magnetic Field Laboratory, Florida State University, Tallahassee, Florida 32306

Received June 5, 2008; E-mail: martin.blackledge@ibs.fr; bruschweiler@magnet.fsu.edu

Abstract: Residual dipolar couplings (RDCs) measured for internally rigid molecular fragments provide important information about the relative orientations of these fragments. Dependent on the symmetry of the alignment tensor and the symmetry of the molecular fragment, however, there generally exist more than one solution for the fragment orientation consistent with the measured RDCs. Analytical solutions are presented that describe the complete set of orientations of internally rigid fragments that are consistent with multiple dipolar couplings measured in a single alignment medium that is rhombic. For the first time, it is shown that, for a planar fragment such as the peptide plane, there generally exist 16 different solutions with their analytical expressions presented explicitly. The presence of these solutions is shown to be highly relevant for standard structure determination protocols using RDCs to refine molecular structures. In particular, when using standard protein structure refinement with RDCs that were measured in a single alignment medium as constraints, it is found that often more than one of the peptide plane solutions is physically viable; i.e., despite being consistent with measured RDCs, the local backbone structure can be incorrect. On the basis of experimental and simulated examples, it is rationalized why protein structures that are refined against RDCs measured in a single medium can have lower resolution (precision) than one would expect on the basis of the experimental accuracy of the RDCs. Conditions are discussed under which the correct solution can be identified.

1. Introduction

Protein structure determination by NMR spectroscopy is primarily based on nuclear Overhauser distance constraints (NOEs) where the achievable resolution is often limited by the resolution of local structure. Residual dipolar couplings (RDCs) provide complementary information on protein structure.[1−3] They can be used for structural refinement[4]and, in favorable cases, even allow the determination of a protein structure with few or no NOEs.[5−8] When considering RDCs alone, multiple solutions exist for the overall orientation of a molecule or molecular fragment for a given set of RDCs. This is due to the mathematical form of the magnetic dipole−dipole interaction and its properties under anisotropic averaging.[9] The exact number of solutions depends on the relative orientations of the internuclear vectors considered and on internal symmetries of the molecule or the molecular fragment under investigation. The situation is similar for dipolar couplings in the solid state,[10−13] where "dipolar waves" have recently been utilized to relate dipolar couplings to secondary structural motifs.[14,15]

Their well-defined tensorial properties make RDCs amenable to analytical mathematical treatment.[16−24] For example, Meiler et al.[17] and Skrynnikov and Kay[24] have determined analytical

† Clark University.
‡ Institut de Biologie Structurale Jean-Pierre Ebel.
§ Florida State University.
# Current address: Biogen Idec, 12 Cambridge Center, Cambridge, MA 02142.

(1) Bax, A.; Kontaxis, G.; Tjandra, N. *Nucl. Magn. Reson. Biol. Macromol., B* **2001**, *339*, 127–174.
(2) Blackledge, M. *Prog. Nucl. Magn. Reson. Spectrosc.* **2005**, *46*, 23–61.
(3) Prestegard, J. H.; Bougault, C. M.; Kishore, A. I. *Chem. Rev.* **2004**, *104*, 3519–3540.
(4) Tjandra, N.; Omichinski, J. G.; Gronenborn, A. M.; Clore, G. M.; Bax, A. *Nat. Struct. Biol.* **1997**, *4*, 732–738.
(5) Delaglio, F.; Kontaxis, G.; Bax, A. *J. Am. Chem. Soc.* **2000**, *122*, 2142–2143.
(6) Hus, J. C.; Marion, D.; Blackledge, M. *J. Am. Chem. Soc.* **2001**, *123*, 1541–1542.
(7) Fowler, C. A.; Tian, F.; Al-Hashimi, H. M.; Prestegard, J. H. *J. Mol. Biol.* **2000**, *304*, 447–460.
(8) Mueller, G. A.; Choy, W. Y.; Yang, D. W.; Forman-Kay, J. D.; Venters, R. A.; Kay, L. E. *J. Mol. Biol.* **2000**, *300*, 197–212.
(9) Saupe, A. *Angew. Chem., Int. Ed. Engl.* **1968**, *7*, 97.
(10) Ketchem, R. R.; Hu, W. D.; Cross, T. A. *J. Cell. Biochem.* **1993**, 277–277.
(11) Tycko, R.; Stewart, P. L.; Opella, S. J. *J. Am. Chem. Soc.* **1986**, *108*, 5419–5425.
(12) Opella, S. J.; Stewart, P. L. *Methods Enzymol.* **1989**, *176*, 242–275.
(13) Quine, J. R.; Cross, T. A. *Concepts Magn. Reson.* **2000**, *12*, 71–82.
(14) Mascioni, A.; Veglia, G. *J. Am. Chem. Soc.* **2003**, *125*, 12520–12526.
(15) Mesleh, M. F.; Veglia, G.; DeSilva, T. M.; Marassi, F. M.; Opella, S. J. *J. Am. Chem. Soc.* **2002**, *124*, 4206–4207.
(16) Meiler, J.; Prompers, J. J.; Peti, W.; Griesinger, C.; Brüschweiler, R. *J. Am. Chem. Soc.* **2001**, *123*, 6098–6107.

expressions for lower and upper bounds for the angle between two dipolar vectors. Wedemeyer et al.[22] developed exact solutions for vector orientations in the presence of additional angular constraints, such as bond angle constraints. Wang and Donald[21] derived analytical expressions for vector orientations when RDCs are available in two alignment media.

When the rhombicity is zero, rotation about the symmetry axis of the alignment tensor does not change the RDCs, and therefore the number of orientations that fulfill a given experimental RDC set belonging to a single alignment is infinite. On the other hand, for nonzero rhombicity, it is commonly assumed that, for a rigid fragment with three or more coplanar dipolar vectors, for example, vectors that lie within the same peptide plane, there are generally eight different solutions for the fragment orientations (see, e.g., refs 6 and 8). It is shown here that, in this case, the number of distinct solutions is in fact twice as large, and analytical expressions are provided that describe all solutions and their orientations. Using the sulfite reductase flavodoxin-like domain as an example, it is demonstrated that multiple solutions can be present within a representative conformational ensemble, even in the case where extensive experimental NOE-derived distance constraints are combined with RDCs from a single data set. In view of the prevalence of this commonly encountered combination of experimental constraints, we discuss in detail the implications for standard structure determination protocols.

2. Theory

We consider a set of interatomic vectors and their associated magnetic dipole−dipole interactions of a molecule expressed in the alignment frame that belongs to an alignment medium with magnitude D_a and rhombicity R. For a rigid internuclear vector $\mathbf{r}_i$ of length r_i, with its direction given by the spherical coordinates (θ_i,ϕ_i), the relationship for a residual dipolar coupling D_i measured for this vector is (in units of Hz)

$$D_i = D_a\left(3\cos^2\theta_i - 1 + \frac{3}{2}R\sin^2\theta_i\cos 2\varphi_i\right) = D_a\left(3z_i^2 - 1 + \frac{3}{2}R(x_i^2 - y_i^2)\right) \quad (1)$$

where $D_a = -\mu_0\gamma_k\gamma_l h/(16\pi^3 r_i^3)A_a$, R is the rhombicity with $0 \le R \le {}^2/_3$, (x_i,y_i,z_i) are the Cartesian coordinates of $\mathbf{r}_i/r_i$ (i.e., after normalizing $\mathbf{r}_i$), A_a is the unitless absolute alignment magnitude, h is Planck's constant, μ_0 is the vacuum permeability, and γ_k, γ_l are the gyromagnetic ratios of the two interacting spins.

Enumeration of Solutions. In general, there are an infinite number of vectors that correspond to a particular value of D_i. One way to reduce this degeneracy is to consider a structural motif with multiple dipolar vectors with known relative orientations. Even in this case, multiple distinguishable orientations of the motif exist that are consistent with the RDCs. A convenient way to enumerate the solutions uses the right-hand-side of eq 1, which makes it immediately clear that the sign inversion of either the x, y, or z axis leaves D_i invariant; i.e., if (x,y,z) is a solution, then any of the eight sign combinations $\mathbf{e}_i = (\pm x,\pm y,\pm z)$ is a solution too.

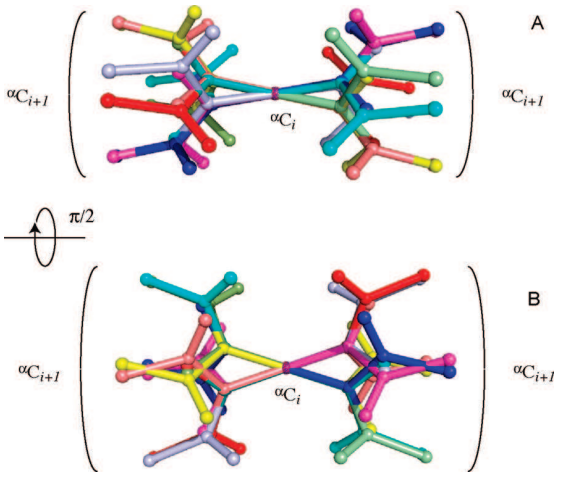

Figure 1. Example of 16 orientations of a peptide plane that are in agreement with a single set of residual dipolar couplings. Cartesian coordinates associated with this example are given in the Supporting Information (Table S1), along with associated RDC values.

For a set of vectors that define a chiral center, however, inversion of the sign with respect to one axis only, for example of all x-components, will change the chirality of the motif and is therefore not allowed. The same applies to the situation where all three axes simultaneously change sign. On the other hand, when two of the three axes change sign, the handedness of the chiral center is conserved. Thus, for a general chiral motif there are only four solutions, (x,y,z), $(x,-y,-z)$, $(-x,y,-z)$, and $(-x,-y,z)$, with the same chirality. (As shown below, for special chiral motives there may be more than four solutions.)

For a structural motif that has no chiral center, which is the case when all vectors lie in the same plane, there is no such restriction, and all eight sign combinations yield allowed solutions. Interestingly, generally there exist *eight additional solutions* that lie in the same planes as the solutions of the first set but whose vector components have orientations that differ from those of the first eight solutions, which is illustrated in Figure 1. These additional solutions flip the planes upside-down (e.g., by a 180° rotation about an axis that lies in the plane). Therefore, in the absence of any in-plane symmetry, the new solutions cannot be aligned with respect to the old solutions by rotation about an axis orthogonal to the plane.

These alternative solutions can be constructed as follows. We consider a plane P depicted in Figure 2A with in-plane vector orientations specified by the three angles (η,ψ,χ) defined in the figure that can be expressed using standard trigonometry:

$$\mathbf{e}(\chi) = (\cos\eta\cos\chi - \sin\eta\sin\psi\sin\chi,\ \sin\eta\cos\chi + \cos\eta\sin\psi\sin\chi,\ \cos\psi\sin\chi) \quad (2)$$

where χ defines the vector orientation within plane P, η is the angle between the intersection of plane P with the xy plane of the alignment frame and the x axis, and ψ is the tilt angle of plane P with respect to the z-axis of the alignment frame. The dipolar couplings can then be expressed as a function of (η,ψ,χ) by insertion of eq 2 into eq 1:

(17) Meiler, J.; Blomberg, N.; Nilges, M.; Griesinger, C. *J. Biomol. NMR* **2000**, *16*, 245–252.
(18) Moltke, S.; Grzesiek, S. *J. Biomol. NMR* **1999**, *15*, 77–82.
(19) Wang, J.; Walsh, J. D.; Kuszewski, J.; Wang, Y. X. *J. Magn. Reson.* **2007**, *189*, 90–103.
(20) Walsh, J. D.; Wang, Y. X. *J. Magn. Reson.* **2005**, *174*, 152–162.
(21) Wang, L. C.; Donald, B. R. *J. Biomol. NMR* **2004**, *29*, 223–242.
(22) Wedemeyer, W. J.; Rohl, C. A.; Scheraga, H. A. *J. Biomol. NMR* **2002**, *22*, 137–151.
(23) Tian, F.; Valafar, H.; Prestegard, J. H. *J. Am. Chem. Soc.* **2001**, *123*, 11791–11796.
(24) Skrynnikov, N. R.; Kay, L. E. *J. Biomol. NMR* **2000**, *18*, 239–252.

$$d' = D/D_a = 3\sin^2\chi\cos^2\psi - 1 + \frac{3}{2}R[(\cos^2\chi(1+\sin^2\psi) - \sin^2\psi)\cos 2\eta - \sin(2\chi)\sin(2\eta)\sin\psi] \quad (3a)$$

or

$$d'(\chi) = D/D_a = A\cos 2\chi - B\sin 2\chi + C \quad (3b)$$

where $A = -(3/2)\cos^2\psi + (3/4)R(1+\sin^2\psi)\cos 2\eta$, $B = (3/2)R\sin(2\eta)\sin\psi$, and $C = (3/2)\cos^2\psi - 1 + (3/4)R(1+\sin^2\psi)\cos(2\eta) - (3/2)R\sin^2\psi\cos(2\eta)$

It can be shown that there exists a suitable axis lying in plane P so that a 180° rotation about this axis of the in-plane dipolar vectors yields additional solutions. This is equivalent to demonstrating that a constant offset δ exists that depends on (η,ψ) so that

$$d'(\chi) = d'(-\chi - \delta) \quad (4)$$

Insertion of eq 3 into eq 4 yields

$$A\cos(2\chi) - B\sin(2\chi) + C = A\cos(2\chi + 2\delta) + B\sin(2\chi + 2\delta) + C \quad (5)$$

Equation 5 is indeed fulfilled for arbitrary angles χ if δ fulfills

$$\tan\delta = B/A \quad (6)$$

where the two solutions for $\delta = \arctan(B/A)$ are 180° shifted with respect to each other, as displayed in Figure 2B.

In summary, if for a rhombic alignment tensor ($R \neq 0$) an internally rigid planar structural motif fulfills a given set of RDCs, then there generally exist no fewer than 16 distinct solutions. The 16 solutions cluster into four groups, where the four solutions of each group lie in the same plane, i.e., each group shares the same (parallel or antiparallel) normal axis to the plane. In the alignment frame, each of these axes points toward one of the four upper quadrants. The four solutions of each group cluster into two subgroups, where the solutions in the subgroups are related to each other by a 180° rotation about the normal axis. The subgroups are related to each other by a 180° rotation about the axis that intersects plane P and the *xy* plane of the alignment frame followed by a rotation about the normal by $-\delta$.

It further follows that a chiral motif consisting of dipolar vectors within a plane and dipolar vectors orthogonal to the plane has the same 16 solutions as the planar system without violating the chirality constraint. A systematic description of all solutions in terms of rotations about the three Euler angles is given in the following sections.

Determination of Solutions from Experimental Data. To determine all possible orientations of an internally rigid molecular fragment on the basis of RDC data, we find it convenient to use internal angular coordinates:

$$D_i = 2D_a\sqrt{\frac{4\pi}{5}}\left\{Y_0^2(\theta_i,\varphi_i) + \sqrt{\frac{3}{8}}R(Y_2^2(\theta_i,\varphi_i) + Y_{-2}^2(\theta_i,\varphi_i))\right\} \quad (7)$$

Equation 7 expresses dipolar couplings in terms of normalized spherical harmonics functions of rank 2, $Y_l^2(\theta,\varphi)$, where (θ_i,φ_i) are the polar angles that belong to the corresponding vector $\mathbf{e}_i$ in the alignment tensor frame. For a set of vectors $\mathbf{e}_i$, one can always define a rotation that rotates the vectors from their initial coordinates (θ_i^0,φ_i^0) in a fixed frame of reference to the frame

(25) Zare, R. N. *Angular Momentum*; John Wiley & Sons: New York, 1988.

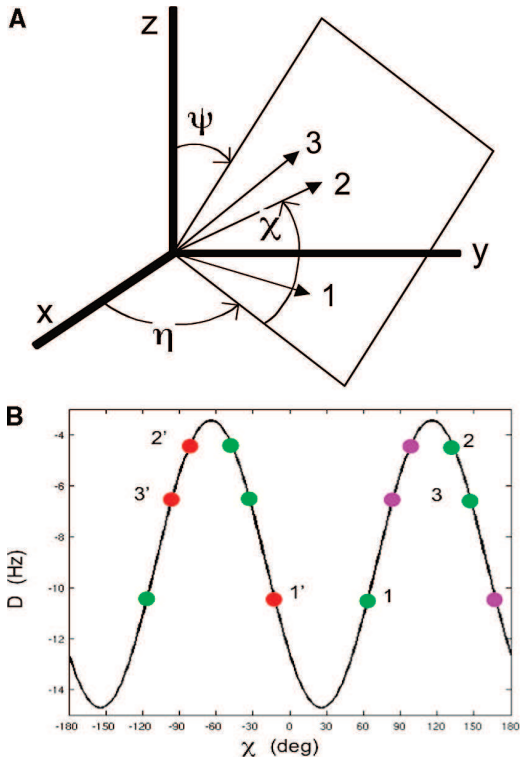

Figure 2. Reverse plane orientations consistent with residual dipolar couplings (RDCs) of three coplanar vectors **1**, **2**, and **3** measured in a single alignment medium with nonzero rhombicity. (A) Definition of angles (η,ψ,χ) that define in-plane vector orientation with respect to alignment frame (x,y,z). (B) For an alignment tensor with $R = 0.35$, the four solutions for vector orientations that lie in the same plane are indicated by four different colors: the red and magenta solutions are related to each other by a $\Delta\chi = 180°$ in-plane rotation. The two sets of solutions (red/magenta vs green) are related to each other by inverting χ followed by a shift δ, i.e., $\chi \rightarrow -\chi - \delta$, where δ is defined by eqs 2−6 as illustrated for vector sets **1**, **2**, **3** and **1′**, **2′**, **3′**.

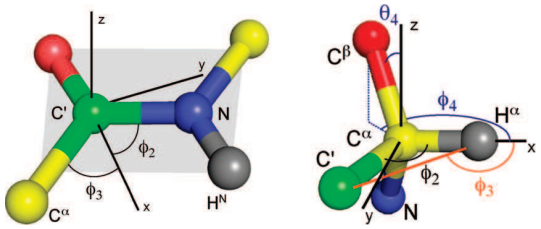

Figure 3. Definition of angles for (A, left) three vectors that lie in the peptide plane and (B, right) four vectors of the chiral amino acid C^α center at the junction between adjacent peptide planes.

of the alignment tensor where their coordinates are (θ_i,φ_i) and their RDC is given by eq 7. Such a rotation can be described by using the Wigner matrices[25] with elements $D^{(2)}_{mn}(\alpha,\beta,\gamma)$ that depend on the Euler angles (α,β,γ). Application of this rotation to eq 7 yields[16]

$$D_i = 2D_a\sqrt{\frac{4\pi}{5}}\left[\sum_{m=-2}^{2} D^2_{m,0}(\alpha,\beta,\gamma)\,Y^2_m(\theta_i^0,\varphi_i^0) + \sqrt{\frac{3}{8}}R\sum_{m=-2}^{2}\left(D^2_{m,2}(\alpha,\beta,\gamma)\,Y^2_m(\theta_i^0,\varphi_i^0) + D^2_{m,-2}(\alpha,\beta,\gamma)\,Y^2_m(\theta_i^0,\varphi_i^0)\right)\right] \quad (8)$$

After simplification, the following expression for the normalized residual dipolar coupling is obtained:

$$d_i = \frac{D_i}{3D_a} = \sum_{k=1}^{5} b_i^k B_k \tag{9a}$$

where B_k are geometric functions that relate the molecular frame to the alignment frame,

$$\begin{aligned}
B_1 &= \cos^2\beta - \frac{1}{3} + \frac{R}{2}\sin^2\beta\cos 2\gamma \\
B_2 &= \cos 2\alpha\left[-\sin^2\beta + \frac{R}{2}(1+\cos^2\beta)\cos 2\gamma\right] - \\
&\quad R\sin 2\alpha \sin 2\gamma \cos\beta \\
B_3 &= \sin 2\alpha\left[-\sin^2\beta + \frac{R}{2}(1+\cos^2\beta)\cos 2\gamma\right] + \\
&\quad R\cos 2\alpha \sin 2\gamma \cos\beta \\
B_4 &= \sin\beta\left[2\cos\beta\cos\alpha\left(1 - \frac{R}{2}\cos 2\gamma\right) + R\sin\alpha\sin 2\gamma\right] \\
B_5 &= \sin\beta\left[2\cos\beta\sin\alpha\left(1 - \frac{R}{2}\cos 2\gamma\right) - R\cos\alpha\sin 2\gamma\right]
\end{aligned} \tag{9b}$$

with (α,β,γ) corresponding to the rotation angles according to the zxz convention (i.e., first rotation by α about the z-axis, second rotation by β about the x-axis, third rotation by γ about the z-axis), and b_i^k represent the dipolar vector orientations in the molecular frame,

$$\begin{aligned}
b_i^1 &= \frac{3\cos^2\theta_i^0 - 1}{2} \\
b_i^2 &= \frac{\sin^2\theta_i^0}{2}\cos 2\varphi_i^0 \\
b_i^3 &= -\frac{\sin^2\theta_i^0}{2}\sin 2\varphi_i^0 \\
b_i^4 &= -\sin\theta_i^0\cos\theta_i^0\cos\varphi_i^0 \\
b_i^5 &= \sin\theta_i^0\cos\theta_i^0\sin\varphi_i^0
\end{aligned} \tag{9c}$$

The linear system of eq 9 can be solved by linear least-squares minimization, e.g., using singular value decomposition (SVD), if there are at least five different dipolar couplings available. This relates to the fact that generally five RDCs are necessary to determine the alignment tensor.[26] Here, we assume that the alignment tensor magnitude D_a and rhombicity R are already known. Therefore, fewer than five RDCs allow establishment of the orientations of a structural motif, as is discussed below.

Equation 9 reflects the fact that, when one knows the coordinates of a set of vectors and their normalized RDCs, one can extract coefficients B_k that contain information about the rotational transformation of these vectors to their correct orientations in the alignment frame. The associated Euler angles and rhombicity R can be extracted from these coefficients as follows:

(a) If $R = 0$:

$$\begin{aligned}
\beta &= \arccos\sqrt{B_1 + \frac{1}{3}} \\
\alpha &= \frac{1}{2}\arctan\frac{B_3}{B_2} = \arctan\frac{B_5}{B_4} \\
\gamma &\quad \text{undefined}
\end{aligned} \tag{10a}$$

which means that there is an infinite number of solutions due to the axial symmetry of the alignment tensor.

(b) If $R \neq 0$:

$$\begin{aligned}
\beta &= \arcsin\sqrt{1 - \frac{\left(B_1 + \frac{4}{3}\right)^2 - B_2^2 - B_3^2}{4 - R^2}} \\
\gamma &= \frac{\pi}{2} - \frac{1}{2}\arccos\frac{2}{R\sin^2\beta}\left(\sin^2\beta + B_1 - \frac{2}{3}\right) \\
\alpha &= \\
&\frac{1}{2}\arctan\frac{\left[-\sin^2\beta + \frac{R}{2}(1+\cos^2\beta)\cos 2\gamma\right]B_3 - R\cos\beta\sin 2\gamma B_2}{\left[-\sin^2\beta + \frac{R}{2}(1+\cos^2\beta)\cos 2\gamma\right]B_2 + R\cos\beta\sin 2\gamma B_3}
\end{aligned} \tag{10b}$$

With the following allowed ranges for the Euler angles, the four-fold degeneracy in orientation, leading to four rotations having the same residual dipolar couplings, is compiled in Table 1:

$$\begin{aligned}
0 \le \alpha \le \pi \\
0 \le \beta \le \frac{\pi}{2} \\
0 \le \gamma \le \pi
\end{aligned} \tag{11}$$

Hence, for the general case of N vectors, eq 10b defines the Euler angles describing the rotation that transforms these vectors from a molecular frame of reference to the alignment frame.

In the presence of symmetries in the molecular fragment under consideration, some of the coefficients B_k can be undetermined, which leads to additional allowed fragment orientations. In the following section, we consider special cases that are relevant for applications to proteins and peptides.

Peptide Plane. The situation where all dipolar vectors lie in the same plane has already been mentioned above. This situation is commonly encountered in practice, with the peptide plane in proteins and nucleic acid bases being the most prominent examples. It is convenient to align the xy plane of the coordinate system with the molecular plane so that all in-plane vectors have $\theta_i^0 = \pi/2$. Here and in the following section, ideal geometries and noise-free RDC data are assumed. As a consequence, in eq 9c, coefficients b_i^4 and b_i^5 are zero, and hence B_4 and B_5 are undetermined, which increases the number of possible solutions. Figure 3A shows the angles that define the peptide plane motif. We use a minimum of three vectors, including $N{-}H^N$, $N^{(i)}{-}C'^{(i-1)}$, and $C'^{(i-1)}{-}C^{\alpha(i-1)}$, whose polar coordinates are as given as follows:

$$\begin{aligned}
&\mathbf{1}\text{:}\quad N{-}H^N,\quad \theta_1^0 = \frac{\pi}{2},\quad \varphi_1^0 = 0 \\
&\mathbf{2}\text{:}\quad C'{-}N,\quad \theta_2^0 = \frac{\pi}{2},\quad \varphi_2^0 = \varphi_2 \\
&\mathbf{3}\text{:}\quad C'{-}C^\alpha,\quad \theta_3^0 = \frac{\pi}{2},\quad \varphi_3^0 = \varphi_3
\end{aligned} \tag{12}$$

Table 1. Four-Fold Degeneracy of Fragment Orientation

solution	Euler angles		
1	α	β	γ
2	α	β	$\gamma + \pi$
3	$\alpha + \pi$	$\pi - \beta$	$-\gamma$
4	$\alpha + \pi$	$\pi - \beta$	$-\gamma + \pi$

(26) Losonczi, J. A.; Andrec, M.; Fischer, M. W. F.; Prestegard, J. H. *J. Magn. Reson.* **1999**, *138*, 334–342.

Equation 9a then leads to

$$d_{N-H^N} = -\frac{1}{2}B_1 + \frac{1}{2}B_2$$
$$d_{C'-N} = -\frac{1}{2}B_1 + \frac{\cos 2\varphi_2}{2}B_2 - \frac{\sin 2\varphi_2}{2}B_3$$
$$d_{C'-C^\alpha} = -\frac{1}{2}B_1 + \frac{\cos 2\varphi_3}{2}B_2 - \frac{\sin 2\varphi_3}{2}B_3 \qquad (13)$$

From the first three equations one obtains for B_1, B_2, and B_3,

$$B_1 = 2\frac{\sin 2\varphi_3(d_{C'-N} - d_{N-H^N}) - \sin 2\varphi_2(d_{C'-C^\alpha} - d_{N-H^N})}{\sin 2\varphi_3(\cos 2\varphi_2 - 1) - \sin 2\varphi_2(\cos 2\varphi_3 - 1)} - 2d_{N-H^N}$$
$$B_2 = 2\frac{\sin 2\varphi_3(d_{C'-N} - d_{N-H^N}) - \sin 2\varphi_2(d_{C'-C^\alpha} - d_{N-H^N})}{\sin 2\varphi_3(\cos 2\varphi_2 - 1) - \sin 2\varphi_2(\cos 2\varphi_3 - 1)}$$
$$B_3 = 2\frac{(\cos 2\varphi_3 - 1)(d_{C'-N} - d_{N-H^N}) - (\cos 2\varphi_2 - 1)(d_{C'-C^\alpha} - d_{N-H^N})}{\sin 2\varphi_3(\cos 2\varphi_2 - 1) - \sin 2\varphi_2(\cos 2\varphi_3 - 1)} \qquad (14)$$

which according to eq 10b leads to Euler angles (α,β,γ). Since in eq 9c coefficients $b_i^4 = b_i^5 = 0$, B_4 and B_5 do not impose any restriction. Therefore, for $R \neq 0$, there are additional solutions (α',β',γ'), which are consistent with the first three equations of eqs 9b and 9c:

$$\alpha' = \frac{1}{2}\arctan\frac{\left[-\sin^2\beta + \frac{R}{2}(1+\cos^2\beta)\cos 2\gamma\right]B_3 + R\cos\beta\sin 2\gamma B_2}{\left[-\sin^2\beta + \frac{R}{2}(1+\cos^2\beta)\cos 2\gamma\right]B_2 - R\cos\beta\sin 2\gamma B_3}$$
$$\beta' = \pi - \beta$$
$$\gamma' = \pi + \gamma \qquad (15)$$

The (α,β,γ) set and the (α',β',γ') set are both eight-fold degenerate, as listed in Table 2. The complete set of peptide plane orientations satisfying the RDCs is then 16, consistent with the Theory section. In special cases, some of these solutions will coincide such that the effective degeneracy can coalesce to ≤8. The two sets of solutions are related to each other by a rotation about an axis that lies in the peptide plane, as discussed above (see also Figures 1 and 2). Figure 4 provides a step-by-step description of Euler angle rotations for the two sets of solutions.

Amino Acid C^α Chiral Center. By using a suitable choice of initial coordinates, eq 9a can be solved analytically in the case of a nonplanar motif. We now demonstrate this result for the C^α chiral center of an amino acid (Figure 3B) using the four one-bond RDCs: $C^\alpha-H^\alpha$, $C^\alpha-C^\beta$, $C^\alpha-C'$, and $C'-H^\alpha$.

Table 2. 16-Fold Degeneracy of Fragment Orientation[a]

solution	Euler angles		
1	α	β	γ
2	α	β	$\gamma + \pi$
3	$\alpha + \pi$	$\pi - \beta$	$-\gamma$
4	$\alpha + \pi$	$\pi - \beta$	$-\gamma + \pi$
5	$\alpha + \pi$	β	γ
6	$\alpha + \pi$	β	$\gamma + \pi$
7	α	$\pi - \beta$	$-\gamma$
8	α	$\pi - \beta$	$-\gamma + \pi$

[a] Solutions 9−16 are identical to 1−8, except for replacing (α,β,γ) by (α',β',γ'), as defined in the text.

We assume the C^α chiral center to have perfect tetrahedral bond geometry, with vector **1** being the $C^\alpha-H^\alpha$ vector along the x axis. The internal coordinates are (Figure 3B)

$$\mathbf{1}: \quad C^\alpha-H^\alpha, \quad \theta_0^1 = \frac{\pi}{2}, \quad \varphi_0^1 = 0$$
$$\mathbf{2}: \quad C^\alpha-C', \quad \theta_0^2 = \frac{\pi}{2}, \quad \varphi_0^2 = \varphi_2$$
$$\mathbf{3}: \quad C'-H^\alpha, \quad \theta_0^3 = \frac{\pi}{2}, \quad \varphi_0^3 = \varphi_3$$
$$\mathbf{4}: \quad C^\alpha-C^\beta, \quad \theta_0^4 = \theta_4, \quad \varphi_0^4 = \varphi_4 \qquad (16)$$

Equation 9 then leads to

$$d_{C^\alpha-H^\alpha} = -\frac{1}{2}B_1 + \frac{1}{2}B_2$$
$$d_{C^\alpha-C'} = -\frac{1}{2}B_1 + \frac{\cos 2\varphi_2}{2}B_2 - \frac{\sin 2\varphi_2}{2}B_3$$
$$d_{C'-H^\alpha} = -\frac{1}{2}B_1 + \frac{\cos 2\varphi_3}{2}B_2 - \frac{\sin 2\varphi_3}{2}B_3$$
$$d_{C^\alpha-C^\beta} = -\frac{3\cos^2\theta_4 - 1}{2}B_1 + \frac{\sin^2\theta_4}{2}\cos 2\varphi_4 B_2 -$$

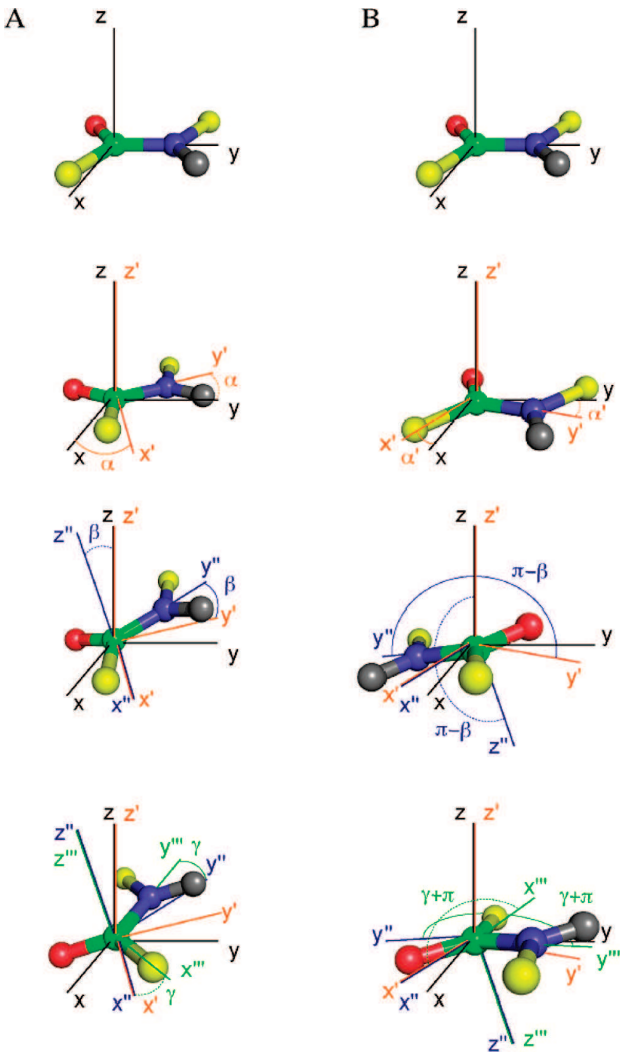

Figure 4. Step-by-step illustration of Euler angle rotations of one vector according to two sets of Euler angles, (α,β,γ) (A, left column) and (α',β',γ') (B, right column).

$$\frac{\sin^2\theta_4}{2}\sin 2\varphi_4 B_3 - \sin\theta_4\cos\theta_4\cos\varphi_4 B_4 + \sin\theta_4\cos\theta_4\sin\varphi_4 B_5$$

One can readily extract B_1, B_2, and B_3 from the $C'-C^\alpha-H^\alpha$ plane using eq 14 and the associated Euler angles (α,β,γ) according to eq 10. Moreover, eq 9b leads to

$$3B_1^2 + B_2^2 + B_3^2 + B_4^2 + B_5^2 = R^2 + \frac{4}{3} \tag{18}$$

with solutions

$$B_4 = \frac{-k_1k_2 \pm \sqrt{k_1^2k_2^2 - (1+k_1^2)\left(3B_1^2 + B_2^2 + B_3^2 - R^2 - \frac{4}{3}\right)}}{1+k_1^2}$$

$$B_5 = k_1B_4 + k_2$$

$$k_1 = \frac{1}{\tan\varphi_4}$$

$$k_2 = \frac{d_{C^\alpha-C^\beta} - \frac{3\cos^2\theta_4 - 1}{2}B_1 + \frac{\sin^2\theta_4}{2}\cos 2\varphi_4 B_2 - \frac{\sin^2\theta_4}{2}\sin 2\varphi_4 B_3}{\sin\theta_4\cos\theta_4\sin\varphi_4} \tag{19}$$

Thus, there are generally two solutions for B_4 and B_5. However, since the Euler angles are known, B_4 and B_5 are also determined by eq 9b and Table 2, which permits identification of the solution that is consistent with the $C'-C^\alpha-H^\alpha$ RDC information. This procedure yields the four sets of unique Euler angles that describe the orientations of the chiral motif that are consistent with the four RDCs. As noticed above, for $\theta_4 = 0$ (vector normal to the plane), the 16 solutions are valid even though it is a chiral motif since $b_4 = 0$, and $b_5 = 0$ and B_4 and B_5 do not enter eq 17.

When the alignment tensor is known, the above formulas allow determination of all fragment orientations consistent with the RDC set. Because these fragments are part of a larger molecular structure to which they are chemically bonded, additional constraints apply such as bond angle and dihedral angle constraints. Moreover, additional information may be available about internuclear distances (from NOEs and trans-hydrogen bond scalar J-couplings) and dihedral angles (from scalar ^3J-couplings and chemical shifts) as discussed in the following section.

3. Application to Biomolecular Structure Determination

By far the most prevalent use of experimental RDCs is found in their application to the refinement of structures whose global architecture is already defined, for example, from distance constraints derived from NOESY experiments. Restrained molecular dynamics calculations are thus routinely used to refine the conformations of proteins and nucleic acids using RDC constraints typically measured in a single alignment medium only. In order to assess the implications of the analytical solutions derived above on commonly applied protocols, we have undertaken a series of calculations using both simulated and experimental data.

Implication of 16 Peptide Plane Orientations for Structural Refinement. Characterization from Simulation. The possible manifestation of the 16 RDC-consistent orientations available for any single planar element using data from one alignment medium has been simulated using RDC-restrained molecular dynamics calculations performed using the program Sculptor[27] under the following conditions:

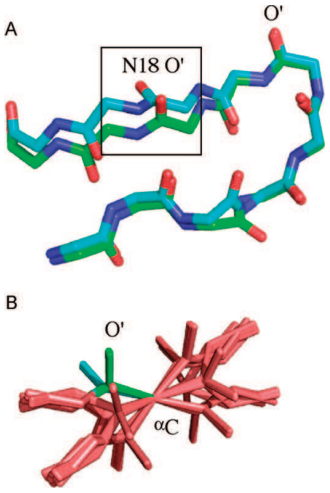

Figure 5. RDC-consistent orientations of peptide plane units produced from RDC-restrained molecular dynamics calculations. (A) Two different RDC-consistent orientations of peptide unit N18-G19 in protein GB3. Data were simulated from a known tensor and subsequently used as restraints, in addition to loose distance restraints between backbone protons (see text). The two structures agree equally well with RDC and distance data as well as tetrahedral geometry. (B) Peptide plane data from peptide plane N18-G19 were used as restraints using the same protocol as in panel A with CNS-Sculptor to determine the possible orientations of the isolated plane. All 16 solutions predicted from eqs 16–18 are shown here. Two of these (the green and the cyan conformations) correspond to the two shown in panel A. Note that four solutions have very similar C^α positions at either end of the plane, illustrating the possibility of finding false minima.

(i) Peptide plane RDCs ($N-H^N$, $C'-C^\alpha$, $N^{(i)}-C'^{(i-1)}$, and $H^N-C'^{(i-1)}$) were simulated from the known structure of protein GB3[28] using a maximally rhombic alignment tensor. Distances were also simulated between all H^N and H^α in the protein, and all distances less than 5 Å were then used as simulated NOE restraints with lower and upper limits of 0 and 6 Å, respectively. The structure was then subjected to a restrained molecular dynamics simulation with only the NOE restraints applied, followed by refinement using both simulated NOEs and simulated RDCs. Both simulations used Cartesian dynamics with a time step of 0.3 fs, with a sampling period of 8 ps at 1000 K, followed by slow cooling to 100 K over 12 ps and subsequent energy minimization. The alignment tensor eigenvalues and eigenvectors were allowed to evolve during the whole period, resulting in accurate reproduction of the original tensor.

Two low-energy structures are shown in Figure 5 that agree equally well with the simulated RDCs. Two planes are seen to have peptide plane orientations that fulfill the RDCs and do not violate tetrahedral geometry nor the loose distance restraints. In order to check whether all these orientations are RDC-consistent as described by eqs 10a,b the RDCs from the N18/G19 plane were used as restraints to determine the orientation of a single, free peptide plane using the known alignment tensor from the simulation. The results of 100 repeated calculations for which RDC energy violations were equivalent and negligible are shown in Figure 5B. The complete set of 16 peptide plane orientations is visibly present, with two of these corresponding to the two orientations shown in Figure 5A. Clearly both

(27) Hus, J. C.; Marion, D.; Blackledge, M. *J. Mol. Biol.* **2000**, *298*, 927–936.
(28) Bouvignies, G.; Markwick, P.; Brüschweiler, R.; Blackledge, M. *J. Am. Chem. Soc.* **2006**, *128*, 15100–15101.

Table 3. Alignment Tensors Used To Simulate RDCs from SiR-FP18 (1−3) and Experimental Tensors from Bacteriophage and Steric Alignment Media (A, B)

	D_a (Hz)	R ($\times 10^{-4}$)	α (°)	β (°)	γ (°)
tensor 1	23	0.65	45.0	0.0	45.0
tensor 2	23	0.65	0.0	45.0	45.0
tensor 3	23	0.65	60.0	60.0	60.0
tensor A	−21.5	0.63	84.6	107.0	117.4
tensor B	−24.2	0.52	−149.8	153.1	−130.2

covalent and nonbonded terms in the force field, as well as additional experimental restraints, will impose additional restrictions on the physical viability of any of these solutions. In particular, orientations that invert the direction of the peptide chain are highly unlikely and are excluded (e.g., on the basis of the position of the subsequent amino acid). We note, however, that four peptide plane orientations have similarly positioned C^{α} atoms, underlining the real possibility that wrong orientations can be found for peptide planes.

(ii) In the second example, experimental distance constraints have been taken[29] from the sulfite reductase flavodoxin-like domain SiR-FP18 (comprising 1114 NOESY-derived distance restraints, 119 hydrogen bond distance restraints based on trans-hydrogen bond $^{3}J_{NC}$ couplings, and accurate dihedral angle restraints from ^{13}C chemical shifts), together with simulated peptide plane RDC data from imposed rhombic alignment tensors (tensors 1 and 2) using a known structure determined using the NOESY dihedral angle and hydrogen-bond distance restraints from randomized initial coordinates.[29] RDCs were also simulated from a third, differently oriented alignment tensor (tensor 3) for purposes of comparison. The use of the third tensor allows for straightforward identification of incorrently oriented planes. These three tensors are linearly independent, as gauged by the scalar product of the tensor elements (Table 3). The same restrained molecular dynamics protocol was applied (see simulation i) to refine the distance-only derived structures. Five calculations were performed with the following active RDCs: (A) no RDCs used in the structure refinement; (B) $N-H^N$ RDCs simulated from tensor 1 used in the structure refinement; (C) four RDCs from the peptide plane ($N-H^N$, $C'-C^{\alpha}$, $N^{(i)}-C'^{(i-1)}$, and $H^N-C'^{(i-1)}$) simulated from tensor 1 used in the structure refinement; (D) four RDCs from the peptide plane ($N-H^N$, $C'-C^{\alpha}$, $N^{(i)}-C'^{(i-1)}$, and $H^N-C'^{(i-1)}$) simulated from tensor 1 and tensor 2 used in the structure refinement; and (E) $N-H^N$ RDCs simulated from tensor 1 and tensor 2 used in the structure refinement. Structural ensembles were selected on the basis of their agreement with distance and RDC restraints.

The final ensemble of RDC-refined structures from calculations B−E are all in close agreement with the respective active RDCs from tensors 1 and 2. These lowest energy (combined experimental energy term associated with RDCs and distance restraints) structures were compared with the RDCs from tensor 3 by fitting the vector orientations of the four RDCs ($N-H^N$, $C'-C^{\alpha}$, $N^{(i)}-C'^{(i-1)}$, and $H^N-C'^{(i-1)}$) to an optimal alignment tensor. This provides a simple means to identify correctly and incorrectly oriented planes. The correlation of simulated and calculated RDCs to the best-fitting tensors is shown in Figure 6 for typical members of the optimal ensembles for calculations A−E.

Importantly, it can be seen that there is no visible improvement between refinement against a single vector from a single alignment medium with respect to using no RDCs. This has previously been noted[30] but is strikingly clear from the correlations in Figure 6A,B. We note here that, using a single set of $N-H^N$ RDCs, it is not always possible to determine the alignment tensor to any accuracy. Alignment tensors are normally determined simultaneously during structure refinement, either using discrete sampling of different combinations of the axial and rhombic components of the tensor or allowing the tensor eigenvalues to float. The latter approach is used here, and while we find that the eigenvalues are sometimes correctly chosen, they can also be wrong, and the target function of the "active" constraints does not reflect this. For example, one of the best structures (on the basis of experimental and physical violations) has an axial tensor component of 31 Hz rather than the value of 23 Hz that was used to simulate the data (this is the worst correlation of the ten structures in the ensemble and is shown in Figure 6B).

In order to investigate this effect in more detail, we have calculated the backbone root-mean-square deviation (rmsd) over ensembles calculated with no RDC information and compared this to the results for ensembles calculated using one full $N-H^N$ RDC set, one full set of four RDCs per peptide plane, and two full sets of peptide plane RDCs. The overall rmsd falls slightly (0.61 compared to 0.68 Å) when $N-H^N$ RDCs are introduced. This information is presented in the Supporting Information on a per-residue basis (Figure S2). In this specific case, local and global structures of SiR FP18 are well defined by the experimental NOE and dihedral angle restraints, as demonstrated by the already low rmsd, and under these conditions, refinement against a single set of $N-H^N$ RDCs provides only minor improvement and does not necessarily better orient the $N-H^N$ vectors compared to the RDC-free refinement. In the light of the high degeneracy of vector orientations that are consistent with a given RDC value, it should not surprise that refinement against a single set of couplings provides negligible improvement. Exceptions are conceivable, however, for example, when the mutual orientations of helical elements are poorly determined due to the sparsity of long-range NOEs. In such cases, the combination of $N-H^N$ vectors from multiple sites may provide a better determination of the relative orientations of the helices.

Beyond a single vector, two vectors (results not shown) and four vectors from a single alignment medium define a planar element (Figure 6C) that reproduces RDCs from tensor 3 for the vast majority of sites and results in more precisely defined backbone conformations (rmsd 0.39 Å). However, the presence of alternative peptide plane orientations was also detected by large local violations of the third set of RDCs (Figure 6C). These are visible as clear outliers in the panels, especially for the $N-H^N$ and $H^N-C'^{(i-1)}$ RDCs. Investigation of these sites reveals that each case corresponds to one of the alternative 15 orientations of the planes described above. Two such cases are shown in Figure 7A,B. For peptide plane D99/Y100, two orientations of the plane can be identified. Both are in agreement with the available distance restraints and tetrahedral geometry at the C^{α} positions. The case of F160/C161 (Figure 7B) illustrates that the orientation of the plane is almost inverted between the two solutions, with the hydrogen−bonding moieties positioned on different sides of the peptide chain. Again no violations of the experimental NOEs are found. The positions of the sites in the protein where two orientations are observed are shown in Figure 7C, illustrating the fact that, while such

(29) Sibille, N.; Blackledge, M.; Brutscher, B.; Coves, J.; Bersch, B. *Biochemistry* **2005**, *44*, 9086–9095.
(30) Grishaev, A.; Bax, A. *Curr. Opin. Struct. Biol.* **2005**, *15*, 563–570.

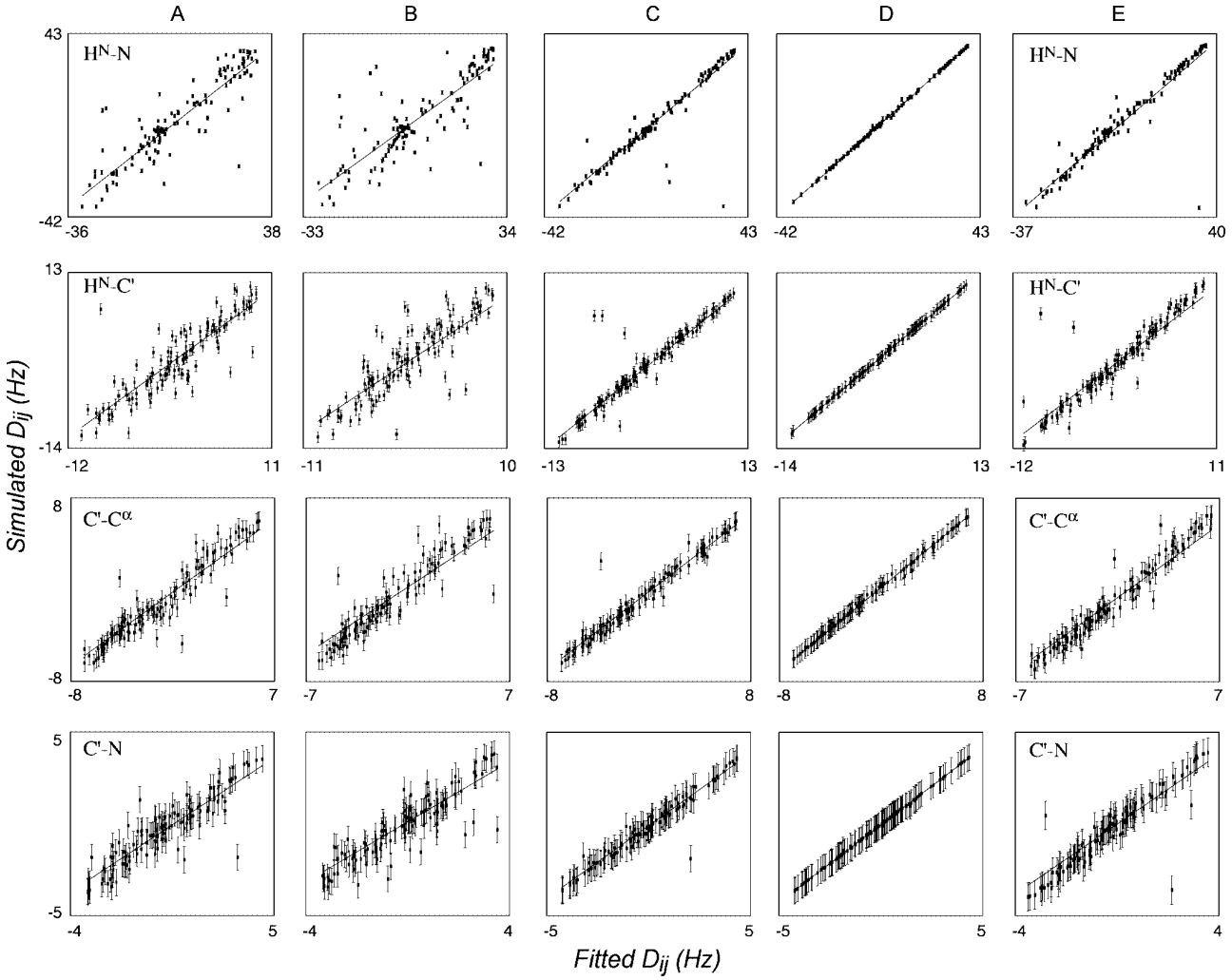

Figure 6. Sampling of RDC-consistent peptide plane orientations using simulated RDCs. Best-fit correlations of RDCs simulated from tensor 3 (Table 3) when fit to optimal alignment tensors determined using structures refined using data from tensors 1 and 2. (A) No RDCs used in the refinement. (B) Only $N-H^N$ RDCs from tensor 1 used in structure refinement. (C) Four peptide plane RDCs from tensor 1 used in structure refinement. (D) Four peptide plane RDCs from tensors 1 and 2 used in structure refinement. (E) Only $N-H^N$ RDCs from tensors 1 and 2 used in structure refinement.

ambiguous peptide plane orientations are more likely to occur in loop regions, due to the generally lower NOE constraint density, they can also occur in regular secondary structural regions. The $C'-C^\alpha$ and $N-C'$ vectors are less sensitive to the presence of incorrect orientations, probably because the overall fold, determined by distance information, essentially defines the position of the C^α atoms, such that the degrees of freedom available to these vectors are more restricted than the vectors involving the H^N atom.

It is also useful to represent the misorientations using a metric that directly relates to the molecular structure. We have therefore calculated the cosines of angles between vectors in each plane and the angle (Ω) between the planes of interest and planes in the target, or known structure. The quantity $\cos(\Omega)$ is shown for each plane for one of the structures present in the ensemble illustrated in Figure 7 (Figure 7D). While for the majority of sites we see that the orientation is correctly determined ($\cos(\Omega) \approx 1$), for other sites the orientation is clearly different, indicating that the plane populates one of the possible equivalent orientations. The cases of orientational degeneracy illustrated in Figure 7A,B are both present and are indicated by asterisks. In the case of plane 160−161, the large angle (>90°) between the correct and incorrect conformations results in a negative value of the cosine.

Not surprisingly, calculations carried out using data from both alignment tensors (Figure 6D) eliminates all of the spurious minima, leaving only the native orientation. The rmsd of the backbone conformation compared to the average structure falls to 0.36 Å, compared to 0.39 Å for a single set of couplings (Figure S2, Supporting Information). Use of RDCs from two differently aligned tensors will indeed reduce the peptide plane orientational degeneracy from 16 to 2, which is the correct plane orientation and its inverse. The structure determination protocol Meccano,[6] developed for the determination of protein backbone conformation using RDCs measured in two different alignment media, exploits the result that the 16-fold degeneracy is reduced to a two-fold degeneracy under these circumstances.

We have also compared the validity of plane orientations refined using only $N-H^N$ RDCs measured in two alignment media (Figure 6E). This again results in isolated occurrences of incorrect orientations of the $N-H^N$ vectors. It is important to note that, in each case, the RDCs used for refinement are essentially in perfect agreement with the structures. Cross-

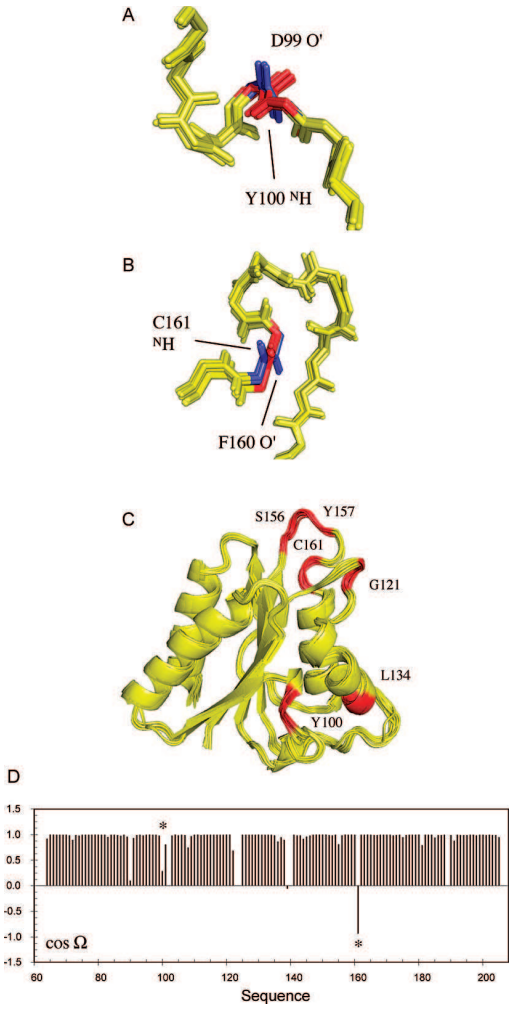

Figure 7. RDC-consistent orientations of peptide plane units produced from RDC-restrained molecular dynamics calculations using experimental distance restraints from sulfite reductase flavodoxin-like domain and simulated RDC data from a single alignment tensor. (A) The 20 structures in the lowest experimental energy ensemble sample two RDC-consistent orientations of peptide unit D99/Y100 that are in agreement with the distance constraints. The blue orientation is the correct orientation. (B) The 20 structures in the lowest experimental energy ensemble sample two RDC-consistent orientations of peptide unit F160/C161 that are in agreement with the distance constraints. The blue orientation again shows the correct orientation. (C) Sites of sulfite reductase flavodoxin-like domain (in red) where two RDC-consistent orientations are observed. (D) Display of cos(Ω) for one of the structures shown in panels A−C, where Ω is the angle between the peptide planes in the structure of interest and the target, or known structure (see text). The positions of peptide planes 99−100 and 160−161 are shown with an asterisk.

validation of the plane orientations using other vectors within the plane does not readily reveal that these orientations are in fact incorrect. It is therefore hard to detect this type of local structural errors in the absence of additional orientational information.

Manifestation in Experimental Ensembles of RDC-Refined Structures. Identification of wrong peptide plane orientations by cross-validation relying on RDCs from different alignments can be directly applied to experimental data. Naturally, demonstration of such effects will also depend on experimental noise in both the "active" data set and the data set used to identify the incorrect equivalent orientations. For this purpose, we have applied the same structure refinement protocols using two high-quality experimental RDC data sets from SiR FP18 in combination with experimental NOESY-derived distances.

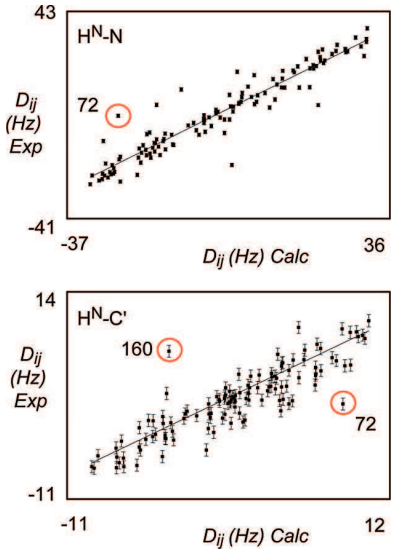

Figure 8. (A) Best-fit correlation of experimental N−H^N RDCs from alignment medium B when fit to optimal alignment tensors determined using structures refined using experimental N−H^N, C′−C^α, and C′−H^N RDCs data from alignment medium A. Correlation from a typical structure is shown from the ensemble best-fitting the active data set (alignment medium A, bacteriophage). A circled outlier corresponds to one of the sites populating two RDC-consistent orientations with respect to these data (see Figure 9). (B) Best-fit correlation of experimental C′−H^N RDCs from alignment medium A when fit to optimal alignment tensors determined using structures refined using experimental N−H^N, C′−C^α, and C′−H^N RDCs data from alignment medium B. Correlation from a typical structure is shown from the ensemble best-fitting the active data set (alignment medium B, alcohol-based alignment). Circled outliers correspond to sites populating two RDC-consistent orientations with respect to these data (see Figure 9).

Alignment medium A (bacteriophage) is expected to align on the basis of electrostatic repulsion, while alignment medium B (PEG/hexanol mixture) is expected to align on the basis of steric repulsion. The resulting alignment tensors have previously been shown to be significantly different (Table 3). The structural ensembles resulting from refinement with respect to media A and B show similar behavior to that observed from the simulated data sets: outliers are observed when compared to the "passive" data sets that are not used in the refinement. In Figure 8 we show correlations with respect to N−H^N and C′−H^N RDCs from typical members of the two ensembles.

Investigation of the structural ensembles reveals that the observed outliers indeed correspond to sites where both correct and incorrect RDC-consistent orientations are sampled. Figure 9 illustrates two such sites. In the case of the plane associating amino acids 158−159, two orientations are found from the refinement with respect to data from medium B, while a single orientation is found in the ensemble refined with respect to data from medium A.

The incorrect orientations in the "B" ensemble violate RDCs from medium A, as shown in Figure 8. In the case of the plane associating amino acids 71−72, two orientations are found from the refinement with respect to data from both media (Figure 9). Of course, only the common orientation is correct, while the incorrect orientations violate RDCs from the unused medium in both cases, as shown in Figure 8. In all cases, the structures in the final ensembles reproduce the active RDCs well. We note that these ensembles are very similar with respect to global rmsd measurements (0.52 ± 0.08 and 0.64 ± 0.07 Å), compared to

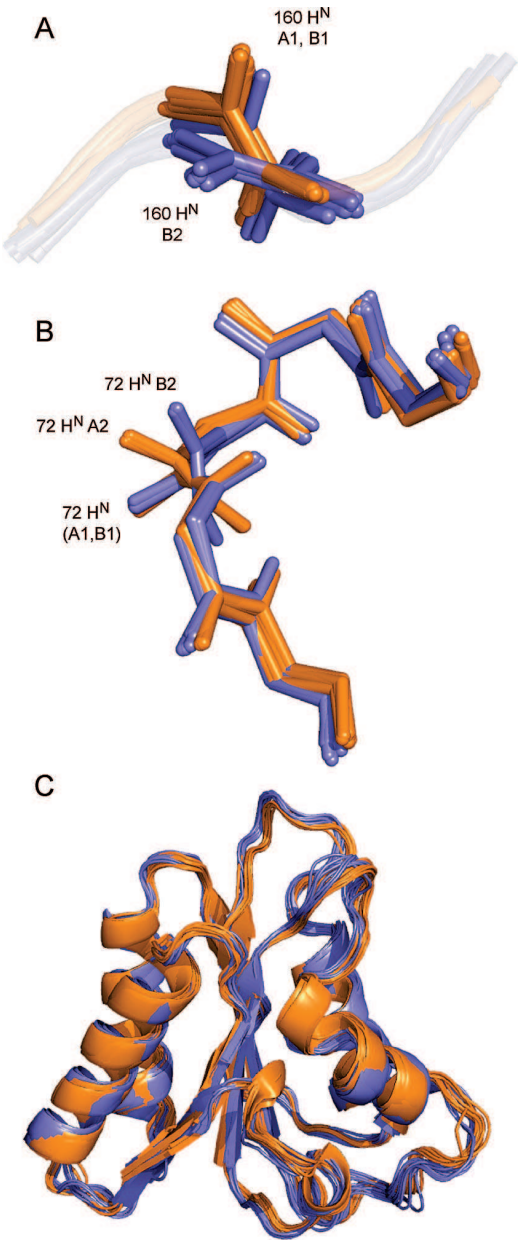

Figure 9. (A) Peptide plane 159−160 shows two orientations in the structural ensemble determined using data from alignment medium B (blue). Two minima, B1 and B2, are indicated. Data from alignment medium A (orange) define a single orientation (A1). (B) Peptide plane 71−72 shows two orientations in the structural ensemble determined using data from alignment medium B (blue: minima B1 and B2). The structural ensemble determined using data from alignment medium A (orange) also show two orientations (A1 and A2). The common orientation is correct. (C) Comparison of the structural ensembles refined using RDCs from media A (orange) and B (blue).

the respective means, and mostly differ in terms of local structure resulting from ambiguous definition of individual peptide plane orientations (Figure 9C).

Implication for Structure Refinement Protocols. Both simulated and experimental data therefore indicate that ambiguous orientations predicted from the analytical descriptions derived above can appear in representative ensembles of structures refined using a single set of RDCs. Their presence is somewhat random, assuming that the correct orientation is equally likely and that other experimental data as well as covalent and noncovalent restraints do not discriminate between the solutions. In this case, the ensemble can contain two or more families of similarly oriented planes. Use of additional RDC restraints from the tetrahedral junction, use of more accurate dihedral angle restraints (e.g., from TALOS), or use of more precise NOE distance information may reduce the possibility of sampling RDC-consistent but wrong peptide plane orientations. In this context it is, however, important to note that the distance information measured from SiR FP18 was relatively complete, including experimentally determined hydrogen-bonding restraints and NOEs extracted from high-resolution NOESY spectra. If one finds numerous sites where multiple orientations are observed under such favorable refinement conditions as illustrated using both noise-free simulated and experimentally measured RDCs, it would be surprising if such cases were not common in standard RDC refinement studies.

A number of additional factors can contribute to the precision of structures refined against RDCs. These include the true geometry of local structural elements that are often constrained to adopt a common optimal geometry. A second source of potential error lies in the supposition that a static model can appropriately describe the experimentally measured parameters. Dynamic averaging, however, can significantly affect measured RDCs[16,31−35] so that refinement against motionally averaged RDCs will be expected to compromise the orientational information present in the final structure. Hence, the analysis presented here primarily applies to amino acids that are part of structured and motionally confined regions of the polypeptide chain.

4. Conclusions

The relationship between a set of RDCs and the associated internuclear vector orientations expressed in an arbitrary molecular frame is determined by the five parameters B_k which are obtained by a linear least-squares fit. Conversion of the B_k parameters to the three Euler angles that define the orientation of the molecular fragment in the alignment frame must take into account that there are multiple solutions due to the symmetry of the alignment tensor and possible symmetries, such as planarity, of the fragment itself. Analytical expressions have been introduced here for the complete set of nonequivalent Euler angles. These relationships, together with bonding constraints, are directly applicable to the reconstruction of biomacromolecular structures from RDCs. Even though this analysis assumes idealized local geometry, it holds interest in methods for the scaffolding of protein backbone structures.

Application of the derived expressions to the description of peptide plane orientations demonstrates that, for nonzero rhombicity, there are generally 16 nonequivalent solutions. Because some of the wrong solutions can have an orientation similar to that of the correct solution, their detection can be difficult,

(31) Tolman, J. R.; Flanagan, J. M.; Kennedy, M. A.; Prestegard, J. H. *Proc. Natl. Acad. Sci. U.S.A* **1995**, *92*, 9279–9283.
(32) Lakomek, N. A.; Walter, K. F.; Fares, C.; Lange, O. F.; de Groot, B. L.; Grubmuller, H.; Brüschweiler, R.; Munk, A.; Becker, S.; Meiler, J.; Griesinger, C. *J. Biomol. NMR* **2008**, *41*, 139–155.
(33) Bernado, P.; Blackledge, M. *J. Am. Chem. Soc.* **2004**, *126*, 4907–4920.
(34) Bouvignies, G.; Bernado, P.; Meier, S.; Cho, K.; Grzesiek, S.; Brüschweiler, R.; Blackledge, M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 13885–13890.
(35) Tolman, J. R.; Ruan, K. *Chem. Rev.* **2006**, *106*, 1720–1736.

directly affecting the validity and resolution of a protein structure determined by standard RDC-assisted refinement. In terms of standard practices for protein structure refinement, we show that the existence of 16 different RDC-consistent peptide plane orientations can result in incorrect local plane orientations, even in the presence of extensive long-range NOE distance information. This has immediate implications for the correctness and resolution of protein and nucleic acid structures that are refined against a single set of RDCs.

Acknowledgment. We dedicate this work to Prof. Richard R. Ernst on the occasion of his 75th birthday. This work was supported by the NIH (grant R01 GM 066041) and NSF (grant 0621482) to R.B. and through EU-NMR JRA3 projects RII3-026145 and ANR NT05-4_42781 to M.B.

Supporting Information Available: Cartesian coordinate representation of 16 distinct solutions for peptide plane orientations; backbone rmsd's of sulfite-reductase flavodoxin-like domain for different RDC refinement protocols. This material is available free of charge via the Internet at http://pubs.acs.org.

JA804274S

Protein Dynamics

DOI: 10.1002/anie.200900476

Protein Conformational Flexibility from Structure-Free Analysis of NMR Dipolar Couplings: Quantitative and Absolute Determination of Backbone Motion in Ubiquitin**

*Loïc Salmon, Guillaume Bouvignies, Phineus Markwick, Nils Lakomek, Scott Showalter, Da-Wei Li, Korvin Walter, Christian Griesinger, Rafael Brüschweiler, and Martin Blackledge**

Molecular dynamics play an essential role in controlling the biological activity of proteins. NMR residual dipolar couplings (RDCs)[1,2] are uniquely sensitive to conformational detail and thus offer a very attractive approach to the characterization of protein dynamics on all time scales up to the millisecond. The simple averaging properties of RDCs make them amenable to rigorous interpretation in terms of protein structure and dynamics. Assumptions made when analyzing protein dynamics from RDCs, as well as the robustness of the resulting dynamic description, can be largely substantiated through self-consistency checks. The recognition of these clear advantages has led to the development of several approaches to the characterization of protein-backbone motions from RDCs, including analytical deconvolution of the amplitudes and anisotropies of bond vectors or structural motifs,[3–8] ensemble-averaging by restrained-molecular-dynamics simulation,[9–11] and direct comparison to unrestrained molecular dynamics (MD).[12,13] In previous studies, however, only relative motional amplitudes could be determined directly from RDCs, because internal motional amplitudes and alignment strength cannot be separated easily. To derive absolute motional amplitudes from RDCs, NMR-relaxation-derived S^2_{Rel} order parameters have been used as upper limits of the corresponding RDC-derived order parameters S^2_{RDC}.[4,10,14]

Herein we describe a robust procedure for the quantitative and absolute determination of protein-backbone motions from RDCs that requires no scaling to an external reference, such as Lipari–Szabo order parameters. We have developed a novel, structure-free approach for the determination of the average orientation of each independent peptide plane in the protein and the associated local conformational dynamics about this mean. Motion is described by using the GAF (Gaussian axial fluctuation)[15–17] model of peptide-plane reorientation, whereby the amplitude, direction, and distribution of peptide-plane motions are analytically described. All alignment-tensor elements are determined simultaneously, which enables the quantitative assessment of the distribution of dynamic amplitudes. The approach was used to describe the conformational dynamics of the protein ubiquitin from experimental RDCs,[10,14,18–20] and the results are compared to motional modes extracted from a long (400 ns) MD simulation.

RDCs subject to conformational averaging in a folded protein can be expressed in terms of orientational bond-vector averages relative to a common molecular alignment frame:

$$D_i^j = -2D_a^j\sqrt{\frac{4\pi}{5}}\left\{\langle Y_0^2(\theta_i,\phi_i)\rangle + \sqrt{\frac{3}{8}}R\left(\langle Y_2^2(\theta_i,\phi_i)\rangle + \langle Y_{-2}^2(\theta_i,\phi_i)\rangle\right)\right\} \quad (1)$$

in which $D_a^j = -A_a^j\left(\mu_0\gamma_k\gamma_l h/16\pi^3 r_i^3\right)$ and R is the rhombicity. When analyzing the spherical harmonic averages $\langle Y_m^2(\theta_i,\phi_i)\rangle$ from RDCs, it is important to accurately determine the alignment tensors (j), and in particular the amplitude term for each tensor A_a^j. If tensors are determined without considering motion, average components of anisotropic and isotropic motional modes are absorbed into A_a^j so that the effective molecular alignment appears lower.[21] As a result, dynamic amplitudes extracted by any of the proposed approaches are reduced. Additional uncertainty may result from the coordinate sets used to estimate alignment tensors.

In this study, we addressed these issues by applying a structure-free GAF approach (SF GAF) to the elucidation of local motions from RDCs. Although in principal the method requires as few as three independent alignment media, in this case $^1D_{NH}$ couplings from 24 alignment media, $^1D_{C'N}$ and $^2D_{C'NH}$ couplings from five alignment media, and $^1D_{C'Ca}$ couplings from two alignment media were used. These data sets were chosen previously from a larger data set on the basis of self-consistency in terms of structure and dynamics.[22,14]

[*] L. Salmon,[+] Dr. G. Bouvignies,[+] Dr. P. Markwick, Dr. M. Blackledge
Protein Dynamics and Flexibility
Institute de Biologie Structurale Jean-Pierre Ebel
CNRS-CEA-UJF UMR 5075
41 rue Jules Horowitz, 38027 Grenoble Cedex (France)
Fax: (+33) 438-789-554
E-mail: martin.blackledge@ibs.fr
Dr. S. Showalter, Dr. D. W. Li, Prof. R. Brüschweiler
Chemical Sciences Laboratory
Department of Chemistry and Biochemistry and
National High Magnetic Field Laboratory (NHMFL)
Florida State University, Tallahassee, FL 32306 (USA)
Dr. N. Lakomek, K. Walter, Prof. C. Griesinger
Department of NMR-Based Structural Biology
Max Planck Institute for Biophysical Chemistry
Am Fassberg 11, 37077 Goettingen (Germany)
[+] These authors contributed equally to this work.

[**] This research was supported by the EU through EU-NMR JRA3 and by the French Research Ministry through ANR-07-PCVI-0013

Supporting information for this article is available on the WWW under http://dx.doi.org/10.1002/anie.200900476.

The following function is minimized for each peptide plane for which a sufficient number of RDCs (>15) are available:

$$\chi^2\left[\{\theta,\phi,\psi\}_i,\{A\}_j,\{S,\sigma_\alpha,\sigma_\beta,\sigma_\gamma\}_i\right]=\sum_{ij}\left(D_{i,j}^{\exp}-D_{i,j}^{\mathrm{calc}}\right)^2/\delta_{i,j}^2 \qquad (2)$$

in which the angles (θ,ϕ,ψ) describe the mean orientation of plane i, $\boldsymbol{A}$ represents the alignment tensor j, σ_α, σ_β ,σ_γ, and S are the amplitudes of the GAF or axially symmetric motions, and δ_{ij} is the estimated weighting of each RDC dataset, as determined by using a robust scaling estimator.[23] Because every peptide plane is treated separately from the others, this approach is "structure-free"; that is, the 3D protein fold is neither required nor constructed, in contrast to previous GAF-based applications.[6,7] The internal geometry of each peptide plane is defined by average heavy-atom coordinates extracted from ultra-high-resolution protein crystal structures.[6] The position of the H^N atom was optimized previously from extensive RDCs measured in protein G, with an optimal average $N{-}H^N$ distance of 1.02 Å,[21] consistent with ^{15}N-relaxation analysis. A more recent analysis suggests a distance of around 1.024 Å.[24] The use of this distance results in essentially identical results (see the Supporting Information). This procedure determines the optimal level of alignment relevant to the dominant motional mode. The result is then refined by cross-validation of "free datasets" by using the full 3D GAF description, which results in slightly higher tensors, probably owing to small-amplitude axially symmetric components that are otherwise absorbed into all A_a^j values. The procedure was applied to RDCs simulated from an MD trajectory of protein G;[25] four known alignment tensors were used, and the procedure was shown to be both valid and accurate to around 1% (see the Supporting Information).

By using the tensors defined in this way, 3D GAF analyses were applied to determine either I $\{\sigma_\gamma\}$, II $\{\sigma_\beta,\sigma_\gamma\}$, or III $\{\sigma_\alpha,\sigma_\beta,\sigma_\gamma\}$. For models I and II, values of $<\sigma_\alpha>=4°$ (I, II) and $<\sigma_\beta>=6°$ (I) were determined by optimizing averages from throughout the molecule. To avoid overfitting, models II and III were only invoked if the increased complexity of the model was statistically justified (on the basis of standard F-tests or AIC);[26] otherwise the average value was retained.

A comparison of the order parameters determined for the $N{-}H^N$ bond vectors by 3D GAF analysis (S^2_{RDC}) with those extracted from the trajectories of a long-timescale (400 ns) MD simulation (S^2_{MD})[12] revealed a similar profile and similar amplitude of motion (Figure 1 A). On average, the nature of the dynamics is very similar, as revealed by a comparison of motional amplitudes about the γ ($^{\alpha}C{-}^{\alpha}C$) axis (Figure 1 B). Only a few values that are close to zero, and are therefore expected to have a large error, show deviations.[25] S^2_{RDC} values were also compared to spin-relaxation order parameters S^2_{Rel}[27] (Figure 1 A). Within experimental uncertainty, three S^2_{RDC} values were found to be significantly higher than S^2_{Rel} (assuming an uncertainty of 0.03 for S^2_{Rel}): a nonphysical situation that may be partly caused by uncertainties in relaxation-data analysis. In previous analyses of the $N{-}H^N$ RDCs by self-consistent RDC-based model-free analysis (SCRM), alignment tensors were determined by using a static model, values were extracted for $\langle Y_m^2(\theta_i,\phi_i)\rangle$, and a

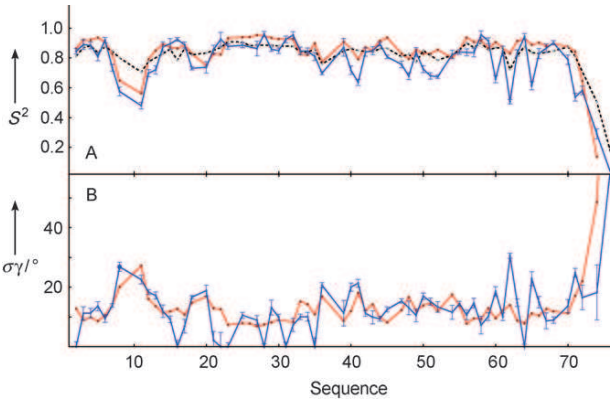

Figure 1. A) $N{-}H^N$ order parameters determined from SF GAF analysis of RDCs for ubiquitin (S^2_{RDC}: blue), 400 ns MD simulation (S^2_{MD}: red), and ^{15}N spin relaxation (S^2_{Rel}: dashed line). B: Amplitude of the γ component of 3D GAF motion (σ_γ) as determined from SF GAF RDC analysis (blue) and 400 ns MD simulation (red). Error bars are derived from noise-based Monte Carlo simulation.

scaling of S^2_{RDC} was applied to fulfill $S^2_{RDC} \le S^2_{Rel}$.[28,14] Applied to the relaxation values shown in Figure 1, this procedure revealed similar order parameters (Figure 2); it reproduced the details of the sequence-dependent dynamics found by using the 3D GAF approach.

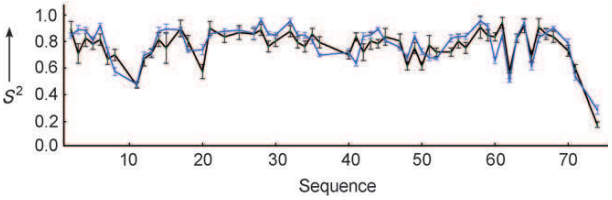

Figure 2. Comparison of $N{-}H^N$ order parameters determined from SF GAF (blue) and SCRM (black) values. The SCRM values were scaled to be equal to or lower than the relaxation data set.[14]

It was instructive to investigate the dependence of these findings on the magnitude of the alignment tensors (Figure 3). An overall scaling factor applied to all tensors was varied over a range $K_A=0.9$–1.1 of the determined A_a^j values. For each point, the local dynamic analysis was repeated with the scaled tensors. The resulting χ^2 value over the entire molecule was calculated for axially symmetric motion (S; model 1), for the best-fitting 1D GAF or axially symmetric motion (1D GAF/S) for each individual plane (model 2), and by 3D GAF analysis (model 3). Model 2 shows a minimum due to anisotropic motion of peptide planes, which dominates the detectable motion, and provides a systematically lower χ^2 value than that obtained with heterogeneously distributed axially symmetric models (if isotropic motion were more appropriate, the curve would be the same as for model 1). The 3D GAF analysis does not show a minimum, as this model can accommodate axially symmetric components while retaining local anisotropy. However, cross-validation of "free datasets" by using the full 3D GAF description does show a minimum,

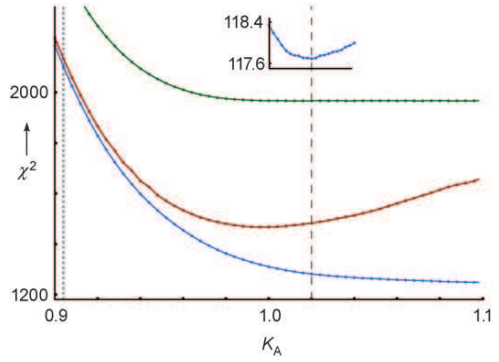

Figure 3. Effects of alignment-tensor scaling on data reproduction with models S (green), 1D GAF/S (red), and 3D GAF (blue). K_A was varied over the range 0.9–1.1 with respect to the minimum determined by 1D GAF/S analysis. The minimum determined from cross-validated 3D GAF analysis (inset) best represents the true alignment tensor.

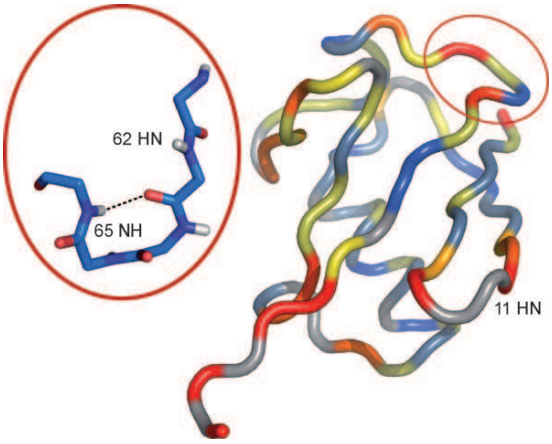

Figure 4. H^N–N order parameters (S^2_{RDC}) from 3D GAF analysis of ubiquitin. Scale from dark blue (1.0) to 0.50 (dark red) via green, yellow, orange. Grey: not determined. Insert: Turn region 62–65 showing higher-amplitude γ/α motions and the hydrogen bond across this turn.

in both experimental and simulated cases (see the inset in Figure 3 and the Supporting Information). The dotted line (on the left-hand side of the graph in Figure 3) shows the alignment tensor determined on the assumption of no dynamics. This alignment tensor is significantly smaller and leads to poorer reproduction of the data.

The robustness of the SF GAF approach was tested by cross-validation: a) Data from each alignment medium were removed for 24 separate analyses, and RDCs were back-calculated by using static and 3D GAF models and appropriate alignment tensors for the specific models; b) two $N–H^N$ RDCs were randomly removed from each peptide plane and back-calculated. In both cases, the χ^2 value was considerably lower for 3D GAF analysis than for the static model (1.1 compared to 4.2 for (a), and 1.0 compared to 3.7 for (b); see the Supporting Information). This demonstrates that the dynamics determined by 3D GAF analysis are required for correct reproduction of the data, providing significant improvement over an analysis using optimal tensors for a static analysis. Average orientations of internuclear bonds determined by this approach are also very similar to those in the high-resolution NMR structure (PDB code: 1d3z;[29] see the Supporting Information).

Enhanced motions apparent in the turn region of the N-terminal β hairpin (8–12) occur in the time range covered by the MD simulation, but are invisible to spin-relaxation experiments due to the overall tumbling of the molecule (around 4 ns). Dynamics occurring in the region of the turn 62–65 appear to reflect larger γ motions for peptide planes 62 and 65 (Figure 4) and are not present in the simulation. A hydrogen bond present in this turn is one of the weakest detected from $^3J_{N,C'}$ scalar coupling.[30] Peptide planes 64 and 65 show the largest-amplitude (30°) α motions in the protein, suggesting mutually dependent dynamics in this turn.

In conclusion, RDCs from ubiquitin measured in multiple alignment media were analyzed in terms of local backbone dynamics by using a structure-free GAF-based approach. This approach yielded significant improvement in data reproduction over a model supposing no dynamics. The method relies only on experimental RDCs; thus, absolute alignment-tensor information and quantitative internal motional modes and amplitudes are obtained from experimental data alone. Comparison with simulation confirmed that the approach can be used to determine internal mobility on an absolute scale. In common with all RDC-based methods, this approach is insensitive to the potential presence of isotropic internal motions of identical magnitude for all dipolar interactions across the whole protein. Nevertheless, there is close correspondence of the average amount, nature, and distribution of motion derived using the SF GAF model compared to the motion present in a 400 ns MD trajectory of ubiquitin. The method also reproduces SCRM-derived order parameters remarkably well despite the fact that no scaling with respect to the relaxation order parameters is applied in the SF GAF approach. More generally, these methods lay the foundation for a quantitative description of the local dynamic behavior of proteins on a wide range of time scales up to the millisecond on the basis of RDCs alone, provided the protein of interest is amenable to study in a sufficient number of different alignment media.[25] This type of analysis provides important information about protein flexibility and will hopefully improve our understanding of the mechanisms governing molecular recognition and function.

Received: January 25, 2009
Published online: May 4, 2009

Keywords: Gaussian axial fluctuation · molecular dynamics · NMR spectroscopy · protein dynamics · residual dipolar coupling

[1] J. R. Tolman, J. M. Flanagan, M. A. Kennedy, J. H. Prestegard, *Nat. Struct. Biol.* **1997**, *4*, 292–297.
[2] N. Tjandra, A. Bax, *Science* **1997**, *278*, 1111–1116.
[3] J. H. Prestegard, H. M. Al-Hashimi, J. R. Tolman, *Q. Rev. Biophys.* **2000**, *33*, 371–424.
[4] J. R. Tolman, *J. Am. Chem. Soc.* **2002**, *124*, 12020–12031.

[5] J. Meiler, J. J. Prompers, W. Peti, C. Griesinger, R. Brüschweiler, *J. Am. Chem. Soc.* **2001**, *123*, 6098 – 6107.
[6] G. Bouvignies, P. Bernado, S. Meier, K. Cho, S. Grzesiek, R. Brüschweiler, M. Blackledge, *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 13885 – 13890.
[7] G. Bouvignies, P. R. L. Markwick, R. Brüschweiler, M. Blackledge, *J. Am. Chem. Soc.* **2006**, *128*, 15100 – 15101.
[8] L. Yao, B. Vogeli, D. Torchia, A. Bax, *J. Phys. Chem. B* **2008**, *112*, 6045 – 6056.
[9] G. M. Clore, C. D. Schwieters, *J. Am. Chem. Soc.* **2004**, *126*, 2923 – 2938.
[10] O. F. Lange, N.-A. Lakomek, C. Farès, G. F. Schröder, K. F. A. Walter, S. Becker, J. Meiler, H. Grubmüller, C. Griesinger, B. L. de Groot, *Science* **2008**, *320*, 1471 – 1475.
[11] K. Lindorff-Larsen, R. B. Best, M. A. DePristo, C. M. Dobson, M. Vendruscolo, *Nature* **2005**, *433*, 128 – 132.
[12] S. A. Showalter, R. Brüschweiler, *J. Am. Chem. Soc.* **2007**, *129*, 4158 – 4159.
[13] P. R. L. Markwick, G. Bouvignies, M. Blackledge, *J. Am. Chem. Soc.* **2007**, *129*, 4724 – 4730.
[14] N. A. Lakomek, K. F. A. Walter, C. Farès, O. F. Lange, B. L. de Groot, H. Grubmüller, R. Brüschweiler, A. Munk, S. Becker, J. Meiler, C. Griesinger, *J. Biomol. NMR* **2008**, *41*, 139 – 155.
[15] R. Brüschweiler, P. E. Wright, *J. Am. Chem. Soc.* **1994**, *116*, 8426 – 8427.
[16] T. Bremi, R. Brüschweiler, *J. Am. Chem. Soc.* **1997**, *119*, 6672 – 6673.
[17] P. R. L. Markwick, R. Sprangers, M. Sattler, *Angew. Chem.* **2005**, *117*, 3296 – 3301; *Angew. Chem. Int. Ed.* **2005**, *44*, 3232 – 3237.
[18] K. B. Briggman, J. R. Tolman, *J. Am. Chem. Soc.* **2003**, *125*, 10164 – 10165.
[19] J. R. Tolman, K. Ruan, *Chem. Rev.* **2006**, *106*, 1720 – 1736.
[20] M. Ottiger, A. Bax, *J. Am. Chem. Soc.* **1998**, *120*, 12334 – 12341.
[21] P. Bernadó, M. Blackledge, *J. Am. Chem. Soc.* **2004**, *126*, 4907 – 4920.
[22] J. C. Hus, R. Brüschweiler, *J. Biomol. NMR* **2002**, *24*, 123 – 132.
[23] P. J. Rousseeuw, C. Croux, *J. Am. Stat. Assoc.* **1993**, *88*, 1273 – 1283.
[24] L. Yao, B. Vögeli, J. Ying, A. Bax, *J. Am. Chem. Soc.* **2008**, *130*, 16518 – 16520.
[25] G. Bouvignies, P. R. L. Markwick, M. Blackledge, *Proteins Struct. Funct. Bioinf.* **2008**, *71*, 353 – 363.
[26] H. Motulsky, A. Christopoulos, *Fitting Models to Biological Data Using Linear and Nonlinear Regression: A Practical Guide to Curve Fitting*, Oxford University Press, Oxford, **2004**.
[27] S. F. Lienin, T. Bremi, B. Brutscher, R. Brüschweiler, R. R. Ernst, *J. Am. Chem. Soc.* **1998**, *120*, 9870 – 9879.
[28] N. A. Lakomek, T. Carlomagno, S. Becker, C. Griesinger, J. Meiler, *J. Biomol. NMR* **2006**, *34*, 101 – 115.
[29] G. Cornilescu, J. L. Marquardt, M. Ottiger, A. Bax, *J. Am. Chem. Soc.* **1998**, *120*, 6836 – 6837.
[30] F. Cordier, S. Grzesiek, *J. Mol. Biol.* **2002**, *317*, 739 – 752.

J|A|C|S
ARTICLES

Published on Web 11/04/2009

Toward a Unified Representation of Protein Structural Dynamics in Solution

Phineus R. L. Markwick,*,†,‡,§ Guillaume Bouvignies,† Loic Salmon,†
J. Andrew McCammon,§ Michael Nilges,‡ and Martin Blackledge*,†

Protein Dynamics and Flexibility, Institute de Biologie Structurale Jean-Pierre Ebel, CNRS-CEA-UJF UMR 5075, 41 rue Jules Horowitz, 38027-Grenoble Cedex, France, Unite de Bio-informatique Structurale, Institut Pasteur, CNRS, URA 2185, F-75015 Paris, France, and Department of Chemistry and Biochemistry, Howard Hughes Medical Institute, University of California San Diego, 9500 Gilman Drive, Urey Hall, La Jolla, California 92003 0365

Received September 3, 2009; E-mail: pmarkwick@ucsd.edu; martin.blackledge@ibs.fr

Abstract: An atomic resolution description of protein flexibility is essential for understanding the role that structural dynamics play in biological processes. Despite the unique dependence of nuclear magnetic resonance (NMR) to motional averaging on different time scales, NMR-based protein structure determination often ignores the presence of dynamics, representing rapidly exchanging conformational equilibria in terms of a single static structure. In this study, we use the rich dynamic information encoded in experimental NMR parameters to develop a molecular and statistical mechanical characterization of the conformational behavior of proteins in solution. Critically, and in contrast to previously proposed techniques, we do not use empirical energy terms to restrain a conformational search, a procedure that can strongly perturb simulated dynamics in a nonpredictable way. Rather, we use accelerated molecular dynamic simulation to gradually increase the level of conformational sampling and to identify the appropriate level of sampling via direct comparison of unrestrained simulation with experimental data. This constraint-free approach thereby provides an atomic resolution free-energy weighted Boltzmann description of protein dynamics occurring on time scales over many orders of magnitude in the protein ubiquitin.

Introduction

Proteins are inherently flexible, displaying a broad range of dynamics over a hierarchy of time-scales from pico-seconds to seconds.[1] This molecular plasticity enables conformational changes in protein backbone and side chains that play critical roles in biomolecular function.[2,3] Nuclear magnetic resonance (NMR) spectroscopy has emerged as the method of choice for studying biomolecular structure and dynamics in solution. All experimentally measured NMR data are affected by motions occurring with characteristic exchange rates that are faster than the so-called chemical-shift range, giving rise to average peaks that represent a potentially complex dynamic average over relatively long time scales (up to the millisecond range for proteins in solution). In addition spin relaxation experiments reflect motions occurring on time scales faster than the molecular rotation diffusion coefficient τ_c (5−20 ns),[4,5] whereas relaxation dispersion can be used to identify sites of slower (μs−ms) conformational exchange.[6−8]

Although the importance of molecular flexibility is generally recognized, standard NMR-based structure determination protocols ignore the presence of protein dynamics, implying that, in common with X-ray crystallography, rapidly exchanging conformational equilibria are routinely represented in terms of a single static structure.[9] The specific averaging properties of different structurally dependent parameters are rarely incorporated into the structure determination procedure, such that the resulting set of coordinates represent a poorly defined average.

The aim of this study is to actively use the rich dynamic information encoded in motionally averaged NMR parameters to develop a structural, dynamic and statistical mechanical molecular representation of the conformational behavior of proteins in solution. Chemical shifts alone are not yet able to describe the dynamics giving rise to the average spectrum, and interproton cross relaxation rates, although rich in structural information, are dependent upon the time scales of the motional processes, and are therefore difficult to interpret quantitatively unless the time scales of the dynamics are known. However, other interactions such as scalar, and possibly more powerfully,

† Institute de Biologie Structurale Jean-Pierre Ebel.
‡ Pasteur Institute.
§ University of California San Diego.

(1) Frauenfelder, H.; Sligar, S. G.; Wolynes, P. G. *Science* **1991**, *254*, 1598–1603.
(2) Karplus, M.; Kuriyan, J. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6679–6685.
(3) Henzler-Wildman, K.; Kern, D. *Nature* **2007**, *450*, 964–972.
(4) Kay, L. E.; Torchia, D. A.; Bax, A. *Biochemistry* **1989**, *28*, 8972–8979.
(5) Palmer, A. G. *Chem. Rev.* **2004**, *104*, 3623–3640.
(6) Mulder, F. A. A.; Mittermaier, A.; Hon, B.; Dahlquist, F. W.; Kay, L. E. *Nat. Struct. Biol.* **2001**, *8*, 932–935.
(7) Palmer, A. G.; Kroenke, C. D.; Loria, J. P. *Methods Enzymol.* **2001**, *339*, 204–238.
(8) Akke, M. *Curr. Opin. Struct. Biol.* **2002**, *12*, 642–647.
(9) Bouvignies, G.; Markwick, P. R.; Blackledge, M. *Chemphyschem* **2007**, *8*, 1901–1909.

10.1021/ja907476w CCC: $40.75

Residual Dipolar Couplings (RDCs), are exquisitely sensitive to conformational detail[10,11] and therefore may hold the key to resolving this long-standing problem.

Over the past decade, RDCs have emerged as powerful tools for studying proteins in solution, providing simultaneous information about time- and ensemble averaged structural and dynamic processes occurring up to millisecond time-scales and thereby encoding key information for understanding biomolecular function.[12,13] Numerous approaches have been proposed to characterize protein backbone conformation from RDCs; most notably the direct determination of dynamic amplitudes and anisotropies of bond vectors or structural motifs from multiple RDC measurements.[14−24] Molecular dynamics (MD) simulation can also provide access to slower motions that can be compared to measured RDCs,[25,26] however despite increasing computational power, trajectories are usually restricted to time-scales of hundreds of nanoseconds, and millisecond trajectories are still not viable. Relatively long simulations (up to 1.2 μs) have identified slow dynamic processes occurring on time-scales beyond the range probed by spin relaxation,[27,28] which would affect the RDC data, but such long simulations provide only a single trajectory in phase space and do not avoid the problem of statistical mechanical sampling. A popular alternative to performing long simulations is to implement time- or ensemble-averaged restraints,[29−31] thereby constraining a multiple copy molecular description to reproduce the conformationally averaged RDCs.[32−34] Although efficient for identifying conformational ensembles in agreement with experimental data, adding an arbitrary pseudopotential to a physical force field can perturb the simulated dynamics in a nonpredictable manner, making further analysis of the resulting trajectories uncertain. More importantly, the generation of an ensemble of structures that can reproduce the experimental data do not necessarily include the relative free energy weighting of each member of the ensemble. The potential energy surface of a protein may be rugged and highly structured, resulting in a broad distribution of populations in conformational space. To accurately reproduce RDCs or any other NMR observable, it would be necessary to include an accurate population analysis.

In this paper, we present a novel approach aimed at providing a self-consistent structural dynamic representation of protein conformational sampling using a combination of state-of-the-art MD simulation and a large set of experimental NMR data.[33,35−38] In contrast to previously proposed techniques, we avoid using empirical energy terms to guide the conformational search, a procedure that has the potential to perturb both the nature and the time scale of simulated dynamics in a nonpredictable way. Rather, we sample conformational space in an unrestrained way, such that Boltzmann statistics can be respected. To sample conformational space efficiently, we use a recently proposed accelerated molecular dynamics (AMD) approach,[39−41] which biases the actual potential energy surface of the protein to enhance transition probabilities between low energy conformational substates. The appropriate level of acceleration, and therefore sampling of conformational space, is directly determined by matching the reproduction of millisecond-averaged experimentally measured dipolar and scalar couplings to those predicted from the different ensembles. The method is used to describe conformational dynamics occurring on time scales over many orders of magnitude in the prototypical protein system ubiquitin and validated against experimental data sensitive to the diverse time scales.

Results

Figure 1 shows averaged order parameters for $^{15}N-^{1}H$ vectors obtained from MD trajectories seeded from AMD simulations using increasing levels of acceleration (see Methods). A heterogeneous distribution of long time-scale dynamics is observed, with increasing amplitude motions occurring predominantly in loop regions (residues 8−11 and 46−48, and to a lesser extent, residues 19, 20, 22, 36 and 60−62). Slower motions are generally seen in regions identified from previously

(10) Tjandra, N.; Bax, A. *Science* **1997**, *278*, 1111–1114.
(11) Hus, J. C.; Salmon, L.; Bouvignies, G.; Lotze, J.; Blackledge, M.; Brüschweiler, R. *J. Am. Chem. Soc.* **2008**, *130*, 15927–15937.
(12) Tolman, J. R.; Flanagan, J. M.; Kennedy, M. A.; Prestegard, J. H. *Nat. Struct. Biol.* **1997**, *4*, 292–297.
(13) Zhang, Q.; Stelzer, A. C.; Fisher, C. K.; Al-Hashimi, H. M. *Nature* **2007**, *450*, 1263–1267.
(14) Tolman, J. R.; Al-Hashimi, H. M. *Annu. Rep. NMR Spectrosc.* **2003**, *51*, 105–166.
(15) Tolman, J. R. *J. Am. Chem. Soc.* **2002**, *124*, 12020–12030.
(16) Meiler, J.; Prompers, J. J.; Peti, W.; Griesinger, C.; Bruschweiler, R. *J. Am. Chem. Soc.* **2001**, *123*, 6098–6107.
(17) Bernado, P.; Blackledge, M. *J. Am. Chem. Soc.* **2004**, *126*, 4907–4920.
(18) Bernado, P.; Blackledge, M. *J. Am. Chem. Soc.* **2004**, *126*, 7760–7761.
(19) Ulmer, T. S.; Ramirez, B. E.; Delaglio, F.; Bax, A. *J. Am. Chem. Soc.* **2003**, *125*, 9179–9191.
(20) Bouvignies, G.; Bernado, P.; Meier, S.; Cho, K.; Grzesiek, S.; Bruschweiler, R.; Blackledge, M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 13885–13890.
(21) Bouvignies, G.; Markwick, P. R. L.; Brüschweiler, R.; Blackledge, M. *J. Am. Chem. Soc.* **2006**, *128*, 15100–15101.
(22) Salmon, L.; Bouvignies, G.; Markwick, P. R. L.; Lakomek, N.; Showalter, S.; Li, D. W.; Walter, K.; Griesinger, C.; Brüschweiler, R.; Blackledge, M. *Angew. Chem., Int. Ed.* **2009**, *48*, 4154–4157.
(23) Tolman, J. R. *Nature* **2009**, *459*, 1063–1064.
(24) Yao, L.; Vogeli, B.; Torchia, D. A.; Bax, A. *J. Phys. Chem. B* **2008**, *112*, 6045–6056.
(25) Showalter, S. A.; Brüschweiler, R. *J. Am. Chem. Soc.* **2007**, *129*, 4158–4159.
(26) Frank, A. T.; Stelzer, A. C.; Al-Hashimi, H. M.; Andricioaei, I. *Nucleic Acids Res.* **2009**, *37*, 3670–3679.
(27) Nederveen, A. J.; Bonvin, A. M. J. J. *J. Chem. Theory Comput.* **2005**, *1*, 363–374.
(28) Maragakis, P.; Lindorff-Larsen, K.; Eastwood, M. P.; Dror, R. O.; Klepeis, J. L.; Arkin, I. T.; Jensen, M.Ø.; Xu, H.; Trbovic, N.; Friesner, R. A.; Iii, A. G.; Shaw, D. E. *J. Phys. Chem. B,* **2008**, *112*, 6155–6158.
(29) Torda, A. T.; Scheek, R. M.; van Gunsteren, W. F. *J. Mol. Biol.* **1990**, *214*, 223–235.
(30) Kemmink, J.; Scheek, R. M. *J. Biomol. NMR* **1995**, *5*, 33–40.
(31) Bonvin, A. M. J. J.; Rullman, J.; Lamerichs, R.; Boelens, R.; Kaptein, R. *Proteins* **1993**, *15*, 385–400.
(32) Clore, G. M.; Schwieters, C. D. *J. Am. Chem. Soc.* **2004**, *126*, 2923–2938.
(33) Lange, O. F.; Lakomek, N. A.; Fares, C.; Schröder, G. F.; Walter, K. F. A.; Becker, S.; Meiler, J.; Grubmüller, H.; Griesinger, C.; De Groot, B. L. *Science* **2008**, *320*, 1471–1475.
(34) Lindorff-Larsen, K.; Best, R. B.; DePristo, M. A.; Dobson, C. M.; Vendruscolo, M. *Nature* **2005**, *433*, 128–132.
(35) Lakomek, N. A.; Walter, K. F. A.; Fares, C.; Lange, O. F.; de Groot, B. L.; Grubmuller, H.; Brüschweiler, R.; Munk, A.; Becker, S.; Meiler, J.; Griesinger, C. *J. Biomol. NMR* **2008**, *41*, 139–155.
(36) Ottiger, M.; Bax, A. *J. Am. Chem. Soc.* **1998**, *120*, 12334–12341.
(37) Ruan, K.; Tolman, J. R. *J. Am. Chem. Soc.* **2005**, *127*, 15032–15033.
(38) Briggman, K. B.; Tolman, J. R. *J. Am. Chem. Soc.* **2003**, *125*, 10164–10165.
(39) Hamelberg, D.; Mongan, J.; McCammon, J. A. *J. Chem. Phys.* **2004**, *120*, 11919–11929.
(40) Hamelberg, D.; McCammon, J. A. *J. Am. Chem. Soc.* **2005**, *127*, 13778–13779.
(41) Markwick, P. R. L.; Bouvignies, G.; Blackledge, M. *J. Am. Chem. Soc.* **2007**, *129*, 4724–4730.

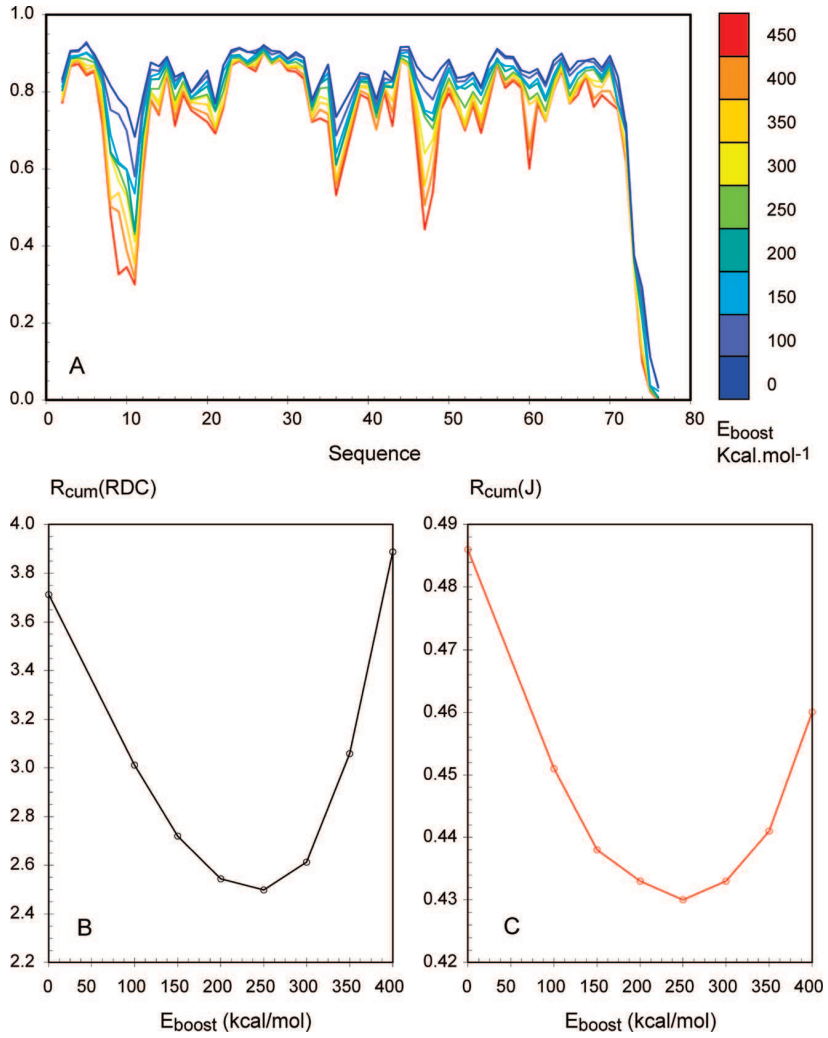

Figure 1. Effect of increasing the acceleration level on N−H^N order parameters in ubiquitin. (A) Order parameters are shown after performing a free energy weighting correction, and are averaged over the trajectories. From top to bottom, the boost energy is 0 (standard 5 ns MD control set), 100, 150, 200, 250, 300, 350, 400, and 450 kcal/mol. The acceleration parameter, α, was fixed at a value of 60 kcal/mol. (B) Change in the trajectory-averaged cumulative R-values for RDCs as a function of the acceleration level. In all cases, the acceleration parameter α = 60 kcal/mol. The boost energy of 0 represents the control set of 5 ns standard MD simulations starting from the X-ray crystal structure[47] using a different random seed generator. (C) Change in the trajectory-averaged cumulative R-values for J-couplings as a function of the acceleration level. In all cases the acceleration parameter α = 60 kcal/mol.

applied long time-scale classical MD simulations of ubiquitin.[25,27,28] A single AMD trajectory, using the equivalent number of steps to 8 ns of classical MD, with relatively low acceleration parameters of (E_{boost} − V_{dih} = 100 kcal mol^{-1}, α = 60 kcal mol^{-1}) results in essentially identical conformational sampling to a 150 ns classical MD trajectory (N−H^N order parameters are compared in Figure S1, Supporting Information). This demonstrates that the AMD approach samples meaningful conformational space analogous to standard long MDs.

1. Agreement between Experimental and Theoretical RDCs and Scalar J-Couplings. To assess the most appropriate level of acceleration, we have analyzed the ability of ensembles derived at each value of (E_{boost} − V_{dih}) to reproduce experimental RDCs and J-couplings (see Methods). Figure 1 shows the trajectory averaged cumulative R-factor (R_{cum}) for RDCs and scalar J-couplings as a function of the acceleration level, and clearly identifies the "optimum" level at (E_{boost} − V_{dih}) = 250 kcal/mol for an acceleration parameter α = 60 kcal/mol. Ensembles generated with less aggressive acceleration sample too little conformational space, while more aggressive acceleration samples too much conformational space to reproduce experimental data.

At the optimal acceleration level, the trajectory-averaged N−H^N RDC cumulative R-factor across all 23 alignment media is 2.496 (average 0.1085). For individual alignment media cumulative R-factors vary between 0.090 and 0.129. Figure 2 shows correlations between experimental and theoretical N−H^N RDCs for four representative data sets with R-factors of 0.096, 0.098, 0.100, and 0.111. Residue specific trajectory-averaged R_{cum} values are compared to a control set calculated from standard 5 ns MD simulations (Figure 3). Only residue 54 shows any increase in the R_{cum} value. Long time-scale dynamics are predominantly located in residues 8−11 and the residue specific R_{cum} values for these residues show improvement compared to the control set. Significant improvement is also observed in residues that show little or no long time-scale dynamics, (e.g.,

(47) Vijay-Kumar, S.; Bugg, C. E.; Cook, W. J. *J. Mol. Biol.* **1987**, *194*, 531–544.

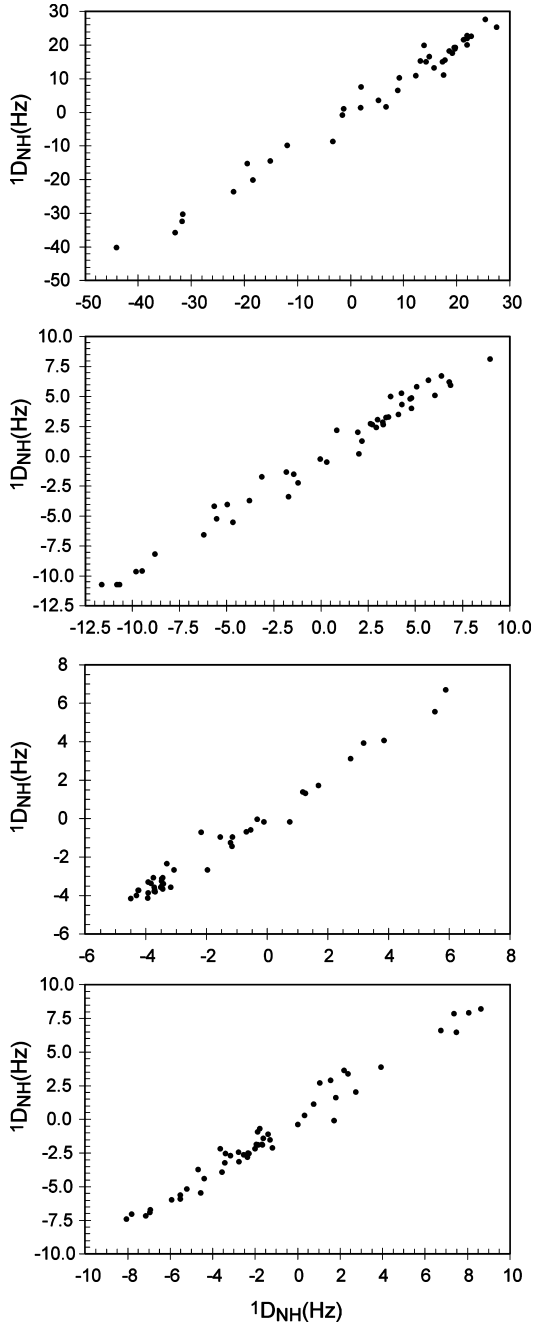

Figure 2. Experimental vs theoretical RDCs for four representative alignment media out of the 23 alignment media. The trajectory averaged cumulative R-factors for the shown alignment media are respectively 0.096, 0.098, 0.100, and 0.111. The trajectory averaged cumulative R-factors across all alignment media varied from 0.090 to 0.129.

15, 17, 34, 42, and 67). The improvement here arises from the more appropriate representation of the time- and ensemble-averaged alignment tensor.

The same level of acceleration also produces the best trajectory-averaged cumulative R-factors for the three scalar J-couplings (H^N-H^α, H^N-C^β and H^N-C'). In comparison to RDCs, scalar J-couplings are less sensitive to the inclusion of long time-scale dynamics, as seen by the relatively small improvement in the R_{cum} values (Figure 2B). This agrees with the previously described phenomenon whereby fitted Karplus

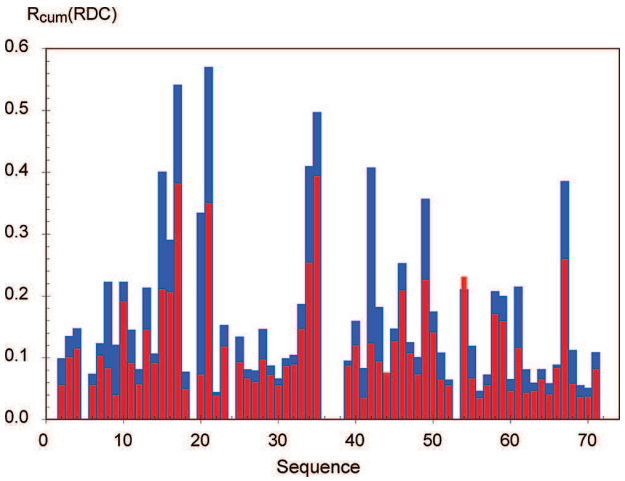

Figure 3. Residue specific trajectory averaged RDC cumulative R-factors. The optimal extended conformational space molecular ensemble [(E_{boost} − V_{dih}) = 250 kcal/mol] is shown in red and compared to a "control set" of standard 5 ns MD simulations shown in blue. Cumulative R-factors for the extended conformational space molecular ensemble are in general lower than those for the control set, confirming the observation of a global improvement in the theoretical RDC data.

parameters can absorb a component of the motion.[42,43] Figure 4 shows the correlation between experimental and theoretical results for the three scalar J-couplings for the molecular ensemble associated with an acceleration level of $E_{boost} - V_{dih} = 250$ kcal/mol. Optimized Karplus curves are compared for $^3J_{NH\text{-}H\alpha}$ obtained from a static structure (1D3Z), standard 5 ns MD simulations and the AMD ensembles and quantum chemistry calculations performed using sum-overstates (SOS) density functional theory (DFT).[44] The curve for the optimal AMD acceleration is almost identical to the DFT-based Karplus curve. A similar effect is observed for the other scalar J-couplings (data not shown).

2. Agreement with Fast Motional Amplitudes Sampled by Spin Relaxation. The free energy weighted molecular ensembles at the RDC-optimum acceleration level provide a representation of the conformational space sampled on time-scales up to the milli-second. These molecular ensembles are composed of individual substates, each with a relative free energy weighting, and from which standard MD simulations have been seeded to probe the fast motions occurring in the local conformational vicinity. Figure 5 depicts the $^{15}N-^1H^N$ order parameters for the fast (ps-ns) dynamics and the effective order parameters probing dynamics on the millisecond time-scale calculated from the free energy weighted ensembles. The fast time-scale $^{15}N-^1H^N$ order parameters obtained by averaging the weighted order-parameters from each substate are in very good agreement with experimental spin relaxation data.[45] In agreement with earlier studies on GB3,[41] we observe an improvement in the agreement between experimental and predicted spin relaxation order parameters when averaging over the extended conformational space ensemble, compared to standard 5 ns MD simulations

(42) Case, D. A.; Scheurer, C.; Brüschweiler, R. *J. Am. Chem. Soc.* **2000**, *122*, 10390–10397.
(43) Markwick, P. R. L.; Showalter, S. A.; Bouvignies, G.; Brüschweiler, R.; Blackledge, M. *J. Biomol. NMR* **2009**, *45*, 17–21.
(44) Malkin, V. G.; Malkina, O. L.; Casida, M. E.; Salahub, D. R. *J. Am. Chem. Soc.* **1994**, *116*, 5898–5908.
(45) Lienin, S. F.; Bremi, T.; Brutscher, B.; Brüschweiler, R.; Ernst, R. R. *J. Am. Chem. Soc.* **1998**, *120*, 9870–9879.

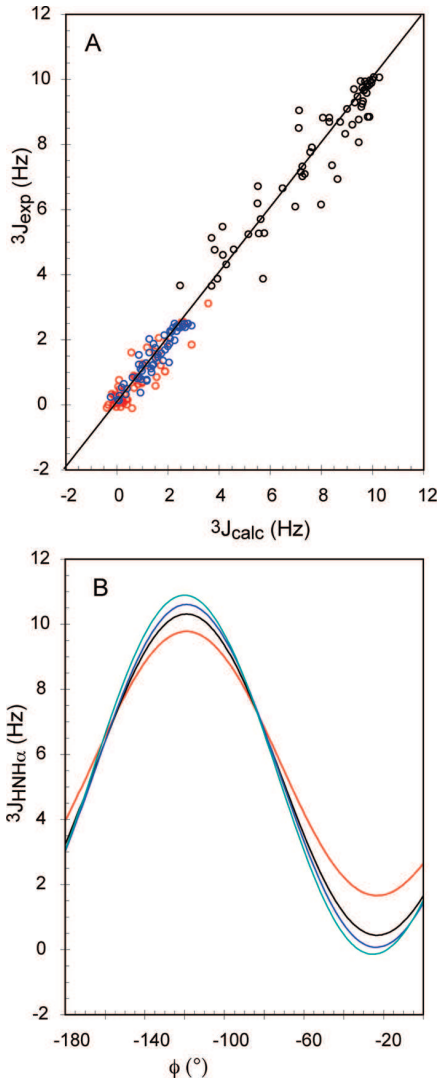

Figure 4. (A) Experimental vs theoretical scalar J-couplings for the optimized extended conformational space molecular ensemble [(E_{boost} − V_{dih}) = 250 kcal/mol]. The three scalar J-couplings are $^3J_{HN\text{-}H\alpha}$ [black circles], $^3J_{HN\text{-}C\beta}$ [red circles], and $^3J_{NH\text{-}C'}$ [blue circles]. (B) NH−H$^\alpha$ Karplus Curves. Red: optimal Karplus curve for ′. Black: optimal Karplus curve for standard 5 ns MD simulation. Blue: optimal Karplus curve for optimal AMD result. Cyan: DFT Karplus curve for NMe-Ala-Ace.

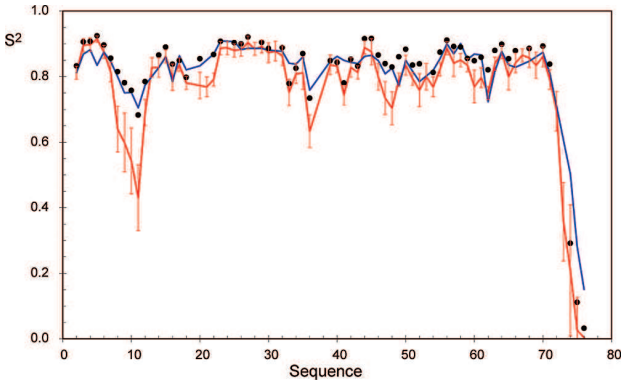

Figure 5. Order parameters for ubiquitin. The ^{15}N spin relaxation experimental data are represented by the blue line.[45] The theoretical fast time-scale (ps−ns) order parameters are shown as black circles and the slow time-scale (RDC-optimized) order parameters are shown in red. The error bars depict the variation in the magnitude of the order parameters for the different molecular ensembles generated from the 20 AMD simulations at the same acceleration level (E_{boost} − V_{dih} = 250 kcal/mol).

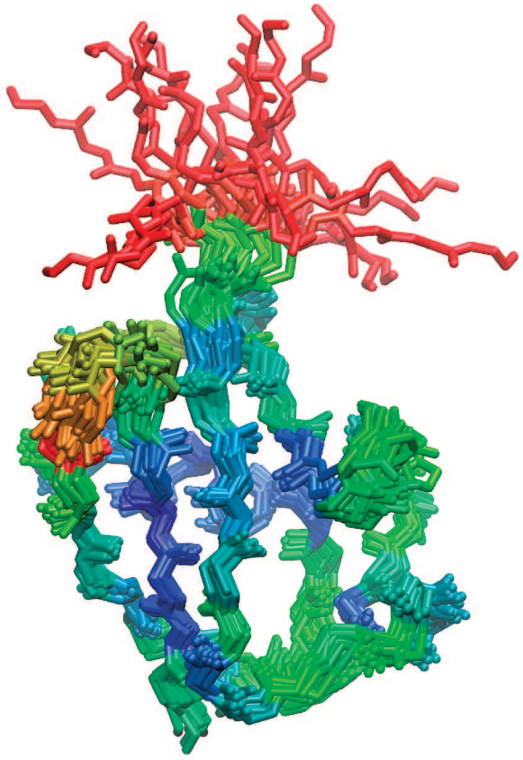

Figure 6. Twenty-four representative structures taken from an RDC-optimized molecular ensemble. The residues are color-coded according to the value of the RDC order parameters (blue: 1.0, red: 0.0).

(data not shown). Figure 6 shows a representative bundle of structures for ubiquitin obtained from the molecular ensemble generated from the RDC-optimal acceleration level.

3. Comparison to Single-Copy and Restrained Ensemble Descriptions. Standard NMR structure refinement against experimental observables generates a time- and ensemble-averaged static conformational representation. In the case of ubiquitin, a high resolution static structure has been optimized (1D3Z),[46] against extensive RDCs, nOes, scalar couplings, and hydrogen bonding restraints. The average R-factor per alignment medium for this structure is 0.093 compared to 0.107 for the optimal AMD ensemble. A direct comparison between unrestrained MD and the NMR structure is complicated by the similarity of the different alignment tensors with those used to refine the structure of 1D3Z. The average RDC R-factor for the optimal AMD ensemble is better than the X-ray crystal structure for ubiquitin (1UBQ)[47] (0.116). Similarly, the AMD approach provides a mean trajectory-averaged J-coupling R-factor (0.143) that reproduces the couplings better than 1UBQ (0.153), and identically to 1D3Z (0.143) despite the fact that these scalar J-couplings were used in the refinement of 1D3Z.

Average backbone coordinates obtained over all free energy-weighted molecular ensembles generated at the RDC-optimal acceleration level are much closer (0.35 Å) to the 1D3Z structure than those obtained from a control set of 5 ns standard MD

(46) Cornilescu, G.; Marquardt, J. L.; Ottiger, M.; Bax, A. *J. Am. Chem. Soc.* **1998**, *120*, 6836–6837.

simulations (0.55 Å). Thus although the RDC-optimal AMD trajectories sample broader conformational space they are distributed about a mean conformation that resembles the time- and ensemble-averaged static structure. Average structures obtained from the optimal molecular ensembles also exhibit negligible violations to the NOE upper- and lower-bounds. Although a quantitative analysis of cross-relaxation rates would require a more rigorous analysis,[48] this demonstrates that the RDC-optimal AMD ensembles are in qualitative agreement with available structural data.

We have also compared the results of the AMD approach to a recent ensemble restrained molecular dynamics (EROS)[33] description of ubiquitin using extensive NOE data and, in this case, all of the 23 RDC data sets treated in our AMD study. Not surprisingly, as the RDCs were used to directly restrain the EROS ensemble, an SVD analysis of the $N-H^N$ RDCs using this ensemble gives a lower average R-factor (0.066).

Principal component projections of the conformational sampling (Supporting Information Figure S2) reveal that the AMD/SVD approach described here samples essentially the same conformational space as the EROS ensemble, without the need to employ ensemble-averaged restraints. This is remarkable considering that no structural restraints are applied, and the conformational space is only defined via global agreement with the entire data set. Closer inspection reveals that the AMD approach further refines the available conformational space sampling following free energy weighting, therefore providing a more realistic structural dynamic representation of the system.

Finally in the recent EROS study, the order parameters derived from the 116 member ensemble were further scaled by a factor of 0.93, on the basis that, in the opinion of the authors, the ensemble may not include a sufficient representation of librational motions. In our case no an additional scaling is applied, so that the AMD ensemble can be considered as a true molecular representation of the dynamic ensemble giving rise to the experimental data. We note that the extent and nature of the dynamics determined using the AMD approach are quantitatively very similar to that determined using the recently developed three-dimensional Gaussian Axial Fluctuation analysis of the experimental RDCs, where, in comparison to spin relaxation derived order parameters, ubiquitin was shown to be essentially rigid on ns-ms time scales, except for the 8−11 hairpin region and some additional surface loops.[22] A more detailed comparison of our results to those obtained from alternative representations, such as 1D3Z and EROS is provided in the Supporting Information.

Discussion

All NMR parameters are affected by motions occurring on time-scales that are faster than the so-called chemical-shift time-scale, resulting in resonance peaks that represent potentially complex dynamic averages over relatively long times (up to the millisecond range for proteins in solution). In this paper, we have combined a novel AMD/SVD approach with extensive experimental NMR data, to provide an accurate description of the structural dynamic behavior of the protein ubiquitin on time-scales ranging from the picosecond to the millisecond. The results of the AMD simulations performed at different acceleration levels confirm that this method can efficiently and accurately sample extended conformational space explored by globular proteins. The SVD analysis allows the model-free determination of the optimal RDC alignment tensor and the optimal J-coupling Karplus parameters, for a given molecular ensemble, obviating the need for calibration against external references or rescaling of order parameters.

The problems of statistical mechanical sampling associated with the incorporation of additional terms into a hybrid potential energy force field, and thereby perturbing the simulated dynamics, are avoided by using restraint-free trajectories seeded at different points of conformational space sampled by the accelerated MD. The accuracy of the resulting RDCs and scalar couplings is however hardly compromised by this procedure, with a similar level of reproduction compared to state-of-the-art single-structure or restrained-ensemble approaches. Importantly fast motional (ps−ns) order parameters derived from experimental spin relaxation data are well reproduced by the population weighted average over MD simulations performed within the different conformational substates. This important result nicely illustrates the potential, inherent to this approach, of resolving the time scales of different motions for comparison with appropriately sensitive experimental data. The optimal AMD molecular ensemble is therefore in agreement with all available experimental data, giving excellent reproduction of RDCs and scalar J-couplings, experimentally determined nOes, as well as ^{15}N spin relaxation data.

Interestingly, the average backbone structure of the optimal molecular ensemble compares very closely with that of the experimentally refined 1D3Z structure, indicating that although these ensembles sample more conformational space, they appear to be distributed about a mean that resembles the experimentally determined time- and ensemble-averaged structure. In all cases, the free energy weighted extended conformational space ensembles reproduce the experimental observables to a substantially greater degree of accuracy than a control set of 5 ns standard MD simulations and provide better reproduction compared to the static X-ray crystal structure (1UBQ).

Conclusions

The ability to provide an explicit description of protein dynamics in terms of conformational substates and associated populations will undoubtedly improve our understanding of the molecular basis of their biological function, and simultaneously provide an essential basis for interpreting dynamically averaged NMR spectra of proteins. A full characterization of protein dynamics requires an integrated experimental and computational approach. In this study we have therefore used enhanced sampling from biased potential molecular dynamics simulation, combined with extensive experimental dipolar and scalar coupling data, to define a self-consistent representative molecular ensemble for solution state protein conformational dynamics. This approach presents a unified structural dynamic representation of the motional properties of proteins in solution that will provide the basis for furthering our understanding of molecular stability, folding, and function, while simultaneously proposing a new methodology for the interpretation of NMR data in terms of molecular ensembles that will be applicable to a wide range of experimental systems.

Methods

Accelerated Molecular Dynamics. The AMD approach involves adding a continuous non-negative bias potential to the potential energy surface of the protein to raise and flatten the potential energy landscape, thereby enhancing the escape rate between low energy

(48) Brüschweiler, R.; Roux, B.; Blackledge, M.; Griesinger, C.; Karplus, M.; Ernst, R. R. *J. Am. Chem. Soc.* **1992**, *114*, 2289–2302.

conformational substates.[39−41] On increasing the level of acceleration, the simulation probes more conformational space. The essential idea behind accelerated molecular dynamics is to define a reference, or "boost energy", E_b, which is fixed above the minimum of the potential energy surface. At each step in the AMD simulation, if the potential energy of the system lies below this boost energy, a continuous, non-negative bias is added to the actual potential. If the potential energy is greater than the boost energy, it remains unaltered. This results in a raising and flattening of the potential energy landscape, decreasing the magnitude of the energy barriers between low energy states, and therefore enhancing the escape rate from one low energy conformational state to another, while maintaining the essential details of the underlying potential energy surface. The extent to which the potential energy surface is modified depends on the difference between the boost energy and the actual potential. Explicitly, the modified potential, $V^*(r)$, is defined as:

$$V^*(\vec{r}) = V(\vec{r}) \quad (1)$$

if the potential energy, V(r), is equal to or greater than the boost energy, and

$$V^*(\vec{r}) = V(\vec{r}) + \Delta V(\vec{r}) \quad (2)$$

if the potential energy is less than the boost energy. The energy modification, or "bias" is given by:

$$\Delta V(\vec{r}) = \frac{(E_b - V(\vec{r}))^2}{\alpha + (E_b - V(\vec{r}))} \quad (3)$$

The extent of acceleration (i.e., how aggressively we enhance the conformational space sampling) is determined by the choice of the boost energy and the acceleration parameter, α. Conformational space sampling can be enhanced by either increasing the boost energy, or decreasing the acceleration parameter. In the present work, the extent of conformational space sampling was controlled by systematically increasing the boost energy using a fixed acceleration parameter. During the course of the simulation, if the potential energy is modified, the forces on the atoms are recalculated for the modified potential. The use of the bias potential defined above ensures that the derivative of the modified potential will not be discontinuous at points where $V(r) = E_b$.

A series of 20, 8 ns, accelerated molecular dynamics (AMD) simulations of ubiquitin were performed at increasing levels of acceleration using the program AMBER8.[49] In each case α was fixed at 60-kcal/mol and the boost energy for the eight acceleration levels was set at 100, 150, 200, 250, 300, 350, 400, and 450-kcal/mol above the dihedral angle energy (estimated from the average dihedral angle energy from the unbiased 5-ns MD simulations).

In all simulations, a time-step of 1 fs and periodic boundary conditions were used with a Langevin thermostat and a Berendsen weak-coupling pressure-stat. Electrostatic interactions were treated using the Particle Mesh Ewald[50] method with a direct space sum limit of 10 Å. The recently developed ff99SB force field was used.[51]

After reweighting the conformational space to obtain the correct canonical Boltzmann distribution, a clustering protocol was implemented to identify low energy conformational substates. A series of short 3 ns standard MD simulations were then seeded from the AMD simulations, to sample the low energy substates. The initial 0.5 ns were discarded, and a MMPB/SA[52] analysis on the resulting MD simulations was used to confirm the AMD free energy weighting protocol. Using these approximate free energies, a set of large (free energy weighted) structural ensembles was generated from the seeded MD simulations for each acceleration level. Resulting ensembles represent free energy weighted trajectories, sampling the conformational space explored by the AMD trajectories at the relevant acceleration level. This method represents an efficient equivalent to performing numerous long time-scale MD simulations. As a control, a series of molecular ensembles were generated from standard 5 ns MD simulations.

The next step is to identify which ensembles can best reproduce the experimental RDC data. At each increasing acceleration level, we have sampled an increasingly large amount of conformational space. Ideally there should exist an optimum sampling on the time-scale relevant to the RDC data. However, for each molecular ensemble, the optimum alignment tensor for a given alignment medium must be calculated. This is achieved in a model free way using a singular value decomposition (SVD) approach,[53,24] and the analysis is performed for available $N-H^N$ RDCs in 23 different alignment media (see details below). Using the optimized alignment tensors, theoretical RDCs were calculated for each molecular ensemble associated with a given acceleration level. As each molecular ensemble represents a single long time-scale trajectory, the theoretical RDCs for each ensemble associated with the same acceleration level were averaged and the agreement between experiment and theory was monitored using the trajectory averaged cumulative R-factor. A similar protocol was performed to calculate scalar J-couplings as outlined below.

Singular Value Decomposition (SVD) and Calculation of RDCs and Scalar J-Couplings. The principal difficulty concerned with the direct calculation of RDCs and scalar J-couplings arises in the determination of parameters defining the strength of the interaction. In the case of RDCs, five unknown parameters are required to explicitly define the alignment tensor. As our simulations are performed in explicit solvent, in the absence of any alignment medium, it is not possible to define explicitly from the simulation alone the preferential alignment of the molecule in a given alignment medium. The issue is further complicated by the fact that the structure, dynamics and preferential alignment of the molecule are mutually dependent: The alignment tensor depends on the shape and anisotropy of the molecule, which is specifically related to the structure. Dynamic motions on different time-scales result in small changes in the shape and anisotropy of the molecule, which, in turn result in small changes in the preferential alignment tensor for a given alignment medium. The approach taken in this work involves the use of an SVD analysis to determine the optimal alignment tensor for each molecular ensemble directly from the experimental data.[53,24] SVD is an exquisite method for solving a set of simultaneous equations. Explicitly, the optimal alignment tensor and RDCs for each molecular ensemble were calculated in a reduced form as:

$$\begin{bmatrix} \langle y^2 - x^2 \rangle & \langle z^2 - x^2 \rangle & \langle 2xy \rangle & \langle 2xz \rangle & \langle 2yz \rangle \\ \langle y^2 - x^2 \rangle & \langle z^2 - x^2 \rangle & \langle 2xy \rangle & \langle 2xz \rangle & \langle 2yz \rangle \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \langle y^2 - x^2 \rangle & \langle z^2 - x^2 \rangle & \langle 2xy \rangle & \langle 2xz \rangle & \langle 2yz \rangle \end{bmatrix} \begin{bmatrix} A_{yy} \\ A_{zz} \\ A_{xy} \\ A_{xz} \\ A_{yz} \end{bmatrix} = \begin{bmatrix} D_{1\text{red}} \\ D_{2\text{red}} \\ \cdots \\ \cdots \\ D_{N\text{red}} \end{bmatrix} \quad (4)$$

where x,y,z are the Cartesian components of the normalized bond vector of interest (in the case of N−H RDCs, the N−H bond vector), A_{ij} is a vector containing the five components necessary to completely define the 3 × 3 alignment tensor (bearing in mind that this tensor is symmetric and traceless) and $D_{i\text{red}}$ is a vector containing the experimental RDCs for the particular alignment medium. The matrix on the left-hand side of the equation, which

(49) Case, D. A.; et al. *AMBER 8*; University of California: San Francisco, CA, 2004.
(50) Cheatham, T. E.; Miller, J. L.; Fox, T.; Darden, T. A.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 4193–4194.
(51) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins: Struct., Funct., Bioinf.* **2006**, *65*, 712–725.
(52) Massova, I.; Kollman, P. A. *J. Am. Chem. Soc.* **1999**, *121*, 8133–8143.
(53) Losonczi, J. A.; Andrec, M.; Fischer, M. W.; Prestegard, J. H. *J. Magn. Reson.* **1999**, *138*, 334–342.

describes the bond vector fluctuations, has dimensions (N,5), where N is the number of RDCs in the given alignment medium. This matrix is formulated from the molecular ensemble where the brackets $\langle .. . \rangle$ represent ensemble averages. SVD of the matrix of bond vector fluctuations liberates the optimal alignment tensor components, from which the theoretical RDCs can be calculated. The principle behind such an analysis is that there should exist some optimal ensemble which represents the conformational space sampled by the system over the time-scales to which RDCs are sensitive (i.e., up to 10-ms for N−H RDCs). For this optimal molecular ensemble, and its optimal SVD-calculated alignment tensor, the resulting theoretical RDCs will be in best agreement with the experimental observables. For molecular ensembles that sample too little or too much conformational space, the SVD analysis will attempt to find the best possible alignment tensor for that particular ensemble, but the resulting RDCs will not be optimal. As mentioned above, we perform a series of AMD simulations at increasing acceleration levels to obtain a set of free energy weighted molecular ensembles that systematically sample an increasing amount of conformational space. By using the SVD analysis to obtain the optimal alignment tensor and hence the theoretical RDCs for each molecular ensemble, we can identify the most appropriate acceleration level (ie. the optimal conformational space sampling) to reproduce the experimental RDCs.

A similar approach was also applied to calculate the backbone scalar J-couplings: We calculated three backbone scalar J-couplings, $^3J(H^N,H\alpha)$, $^3J(H^N,C\beta)$, and $^3J(H^N,C')$. The magnitude of all these J-couplings is strongly related to the backbone ϕ angle and can in general be described using the well-known Karplus equation:[54]

$$^3J(i,j) = A\cos^2(\varphi + \theta) + B\cos(\varphi + \theta) + C \qquad (5)$$

where A, B, and C are the Karplus parameters, and θ is an offset angle, which typically has a value of 180° for $^3J(H^N,C')$, −60° for $^3J(H^N,H\alpha)$ and 60° for $^3J(H^N,C\beta)$. To calculate these scalar J-couplings, we used the SVD analysis to obtain the optimal Karplus parameters for each molecular ensemble:

$$\begin{bmatrix} \langle\cos^2(\varphi_1 + \theta)\rangle & \langle\cos(\varphi_1 + \theta)\rangle & 1 \\ \langle\cos^2(\varphi_2 + \theta)\rangle & \langle\cos(\varphi_2 + \theta\rangle & 1 \\ \ldots & \ldots & \ldots \\ \langle\cos^2(\varphi_N + \theta)\rangle & \langle\cos(\varphi_N + \theta)\rangle & 1 \end{bmatrix} \begin{bmatrix} A \\ B \\ C \end{bmatrix} = \begin{bmatrix} J_1 \\ J_2 \\ \ldots \\ J_N \end{bmatrix} \qquad (6)$$

In each case, the analysis was initially performed using the typical θ offset angles defined above. The θ-offset values were then optimized by changing the θ-offset value in 1° steps and repeating the SVD analysis until the best reproduction of the experimental scalar J-couplings was achieved.

Acknowledgment. This work was supported by the EU through EU-NMR JRA3 and French Research Ministry through ANR Protein-Motion PCV (2007) ANR-07-PCVI-0013 and ANR-06-CIS6-012-01. The work at UCSD was supported in part by NSF, NIH, HHMI, CTBP, and NBCR.

Supporting Information Available: Detailed description of the data analysis and of the comparison of different ensemble descriptions of ubiquitin. Complete ref 49. This material is available free of charge via the Internet at http://pubs.acs.org.

JA907476W

(54) Karplus, M. *J. Chem. Phys.* **1959**, *30*, 11. Karplus, M. *J. Am. Chem. Soc.* **1963**, *85*, 2870–2871.

J|A|C|S
ARTICLES

Published on Web 11/12/2009

Quantitative Description of Backbone Conformational Sampling of Unfolded Proteins at Amino Acid Resolution from NMR Residual Dipolar Couplings

Gabrielle Nodet,[†] Loïc Salmon,[†] Valéry Ozenne,[†] Sebastian Meier,[‡] Malene Ringkjøbing Jensen,[†] and Martin Blackledge*,[†]

Protein Dynamics and Flexibility, Institut de Biologie Structurale Jean-Pierre Ebel, CEA, CNRS, UJF UMR 5075, 41 Rue Jules Horowitz, Grenoble 38027, France, and Carlsberg Laboratory, Gamle Carlsberg Vej 10, 2500 Valby, Denmark

Received August 22, 2009; E-mail: martin.blackledge@ibs.fr

Abstract: An atomic resolution characterization of the structural properties of unfolded proteins that explicitly invokes the highly dynamic nature of the unfolded state will be extremely important for the development of a quantitative understanding of the thermodynamic basis of protein folding and stability. Here we develop a novel approach using residual dipolar couplings (RDCs) from unfolded proteins to determine conformational behavior on an amino acid specific basis. Conformational sampling is described in terms of ensembles of structures selected from a large pool of conformers. We test this approach, using extensive simulation, to determine how well the fitting of RDCs to reduced conformational ensembles containing few copies of the molecule can correctly reproduce the backbone conformational behavior of the protein. Having established approaches that allow accurate mapping of backbone dihedral angle conformational space from RDCs, we apply these methods to obtain an amino acid specific description of ubiquitin denatured in 8 M urea at pH 2.5. Cross-validation of data not employed in the fit verifies that an ensemble size of 200 structures is appropriate to characterize the highly fluctuating backbone. This approach allows us to identify local conformational sampling properties of urea-unfolded ubiquitin, which shows that the backbone sampling of certain types of charged or polar amino acids, in particular threonine, glutamic acid, and arginine, is affected more strongly by urea binding than amino acids with hydrophobic side chains. In general, the approach presented here establishes robust procedures for the study of all denatured and intrinsically disordered states.

Introduction

Despite decades of experimental and theoretical advances in the characterization of structure, kinetics, dynamics, and thermodynamics of many thousands of soluble, folded proteins, the mechanism of protein folding, the conformational transition from a flexible unfolded polypeptide chain to a stable folded protein structure, remains largely unexplained.[1] One reason for this is that one side of the protein folding equation is essentially impossible to characterize in atomic detail using classical approaches to structural biology, requiring instead the development of approaches that explicitly invoke the highly dynamic nature of the unfolded state.[2−5] An atomic-resolution characterization of the structural properties of unfolded proteins is therefore an essential prerequisite for a quantitative understanding of the thermodynamic basis of protein folding and stability.

The importance of developing techniques that are capable of describing the conformational sampling of unfolded polypeptide chains in solution has gained further importance with the gradual realization, over the past decade, that a large fraction of eukaryotic genomes codes for proteins that are intrinsically disordered in their native state.[6−9] Of particular relevance is the relationship between intrinsic structural characteristics of the unfolded chain and the mechanisms of protein folding upon binding, underlining the need for a basic understanding of the conformational space that is populated by a protein in the unfolded state.[10,11] The role that intrinsically disordered proteins (IDPs) play in neurodegenerative disease and cancer further emphasizes the importance of understanding conformational transitions from physiological to pathological forms of the same protein.[12]

Nuclear magnetic resonance (NMR) spectroscopy is probably the most powerful biophysical tool for studying IDPs due to

[†] Institut de Biologie Structurale Jean-Pierre Ebel.
[‡] Carlsberg Laboratory.

(1) Dill, K. A.; Shortle, D. *Annu. Rev. Biochem.* **1991**, *60*, 795–825.
(2) Daggett, V.; Fersht, A. R. *Natl. Rev. Mol. Cell Biol.* **2003**, *4*, 497–502.
(3) Vendruscolo, M.; Paci, E.; Karplus, M.; Dobson, C. M. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 14817–14821.
(4) Mittag, T.; Forman-Kay, J. D. *Curr. Opin. Struct. Biol.* **2007**, *17*, 3–14.
(5) Eliezer, D. *Curr. Opin. Struct. Biol.* **2009**, *19*, 23–30.
(6) Uversky, V. N. *Protein Sci.* **2002**, *11*, 739–756.
(7) Dunker, A. K.; Brown, C. J.; Lawson, J. D.; Iakoucheva, L. M.; Obradovic, Z. *Biochemistry* **2002**, *41*, 6573–6582.
(8) Tompa, P. *TIBS.* **2002**, *27*, 527–533.
(9) Fink, A. L. *Curr. Opin. Struct. Biol.* **2005**, *15*, 35–41.
(10) Dyson, H. J.; Wright, P. E. *Curr. Opin. Struct. Biol.* **2002**, *12*, 54–60.
(11) Fuxreiter, M.; Simon, I.; Friedrich, P.; Tompa, P. *J. Mol. Biol.* **2004**, *338*, 1015–1026.
(12) Dobson, C. M. *Trends Biol. Sci.* **1999**, *24*, 329–332.

10.1021/ja9069024 CCC: $40.75

the remarkable sensitivity of different NMR phenomena to dynamics occurring on time scales varying from picoseconds to hours and the ability to report on both local and long-range structure.[13] In particular, residual dipolar couplings (RDCs), which become measurable when a protein is dissolved in an anisotropic alignment medium or matrix,[14,15] have been shown to be very sensitive reporters of local and long-range structure,[16] even in highly disordered systems.[17] Since the initial demonstration that RDCs can be measured in proteins even under highly denaturing conditions,[18−25] it has been recognized that RDCs provide unique site-specific probes of orientational order in disordered states.[17,26]

A recently developed explicit ensemble description of IDPs, flexible-meccano,[27] constructs multiple copies of the protein in different states, designed to represent all possible conformational states that exchange on time scales relevant to the NMR time scale. Using a statistical coil description that samples amino acid-specific backbone dihedral angle $\{\phi/\psi\}$ propensities, a conformational ensemble is created, and RDCs are calculated for each conformer and then averaged over the ensemble. This approach implicitly assumes that all conformers are in rapid exchange on time scales faster than a millisecond, an assumption based on the presence of a single set of NMR signals detected in ^{1}H and ^{15}N spectra of denatured and intrinsically disordered proteins. The absence of conformational exchange broadening excludes the presence of exchange between significantly populated conformational states occurring on slower time scales. RDCs simulated using these approaches present reasonable agreement with experimental couplings measured in both intrinsically disordered and chemically denatured proteins.[28−32] These studies have been used to provide evidence that site-specific differences in RDCs measured along the primary chain can result from native differences in the rigidity of different amino acid types in an otherwise fully disordered chain,[27] from the presence of transiently populated local secondary structural elements[31] or from the presence of transient interactions between sites distant in the chain.[28]

While ^{15}N−^{1}H^{N} RDCs are by far the most commonly measured dipolar couplings, for reasons of experimental facility and precision, the advantages of measuring more RDCs from different spin-pairs in the peptide unit were recently demonstrated by Meier et al., who determined up to seven RDCs per amino acid from urea-unfolded ubiquitin at pH 2.5, including ^{15}N−^{1}H^{N}, ^{13}C$^{\alpha}$−^{1}H$^{\alpha}$, and ^{13}C$^{\alpha}$−^{13}C′ RDCs, inter- and intraresidue ^{1}H^{N}−^{1}H$^{\alpha}$ RDCs, and ^{1}H^{N}−^{1}H^{N} RDCs measured using quantitative *J*-type experiments[33] on perdeuterated ubiquitin. In combination, these data indicated that the standard description of the statistical coil behavior was inappropriate for urea unfolded proteins and that a modification of the random coil description was necessary to account simultaneously for all data.[34] On the basis of extensive simulation, the authors proposed that, in the presence of urea, the backbone dihedral angles defining the conformational behavior of the unfolded chain have a significantly higher propensity to sample more extended regions of Ramachandran space ($\psi > 50°$, $\phi < 0°$). This indication is supported by a comparison of extensive experimental small angle scattering (SAS) and pulse field gradient (PFG) dependences measured from urea-denatured proteins, with predicted data from conformational ensembles constructed using statistical coil models sampling increasing levels of this extended region (P. Bernado, personal communication). These independent biophysical techniques concur to substantiate an overall description of conformational bias respected by disordered polypeptide chains in the presence of high concentrations of denaturant.[35−38] RDCs measured between different spins within the peptide unit have also been shown to exhibit complementary dependences on the presence of local structure, an observation that has been shown to be crucial for the quantitative determination of the nature and extent of helical sampling present in molecular recognition elements of intrinsically disordered viral proteins[31] and the disordered N-terminal domain of p53.[39]

These studies have mainly used a rational, hypothesis-based approach, calculating explicit ensembles containing tens of thousands of conformers from different conformational sampling regimes and comparing the ensemble-averaged couplings to experimental data. In this study, we are interested in taking the analysis of RDCs one crucial step further, by investigating the possibility of defining the conformational sampling of the peptide chain directly from the experimental NMR data at amino

(13) Dyson, H. J.; Wright, P. E. *Chem. Rev.* **2004**, *104*, 3607–3622.
(14) Tjandra, N.; Bax, A. *Science* **1997**, *278*, 1111–1114.
(15) Prestegard, J. H.; al-Hashimi, H. M.; Tolman, J. R. *Q. Rev. Biophys.* **2000**, *33*, 371–424.
(16) Blackledge, M. *Prog. Nucl. Magn. Reson. Spectrosc.* **2005**, *46*, 23–61.
(17) Meier, S.; Blackledge, M.; Grzesiek, S. *J. Chem. Phys.* **2008**, *128*, 052204.
(18) Shortle, D.; Ackerman, M. S. *Science* **2001**, *293*, 487–489.
(19) Alexandrescu, A. T.; Kammerer, R. A. *Protein Sci.* **2003**, *12*, 2132–2140.
(20) Mohana-Borges, R.; Goto, N. K.; Kroon, G. J. A.; Dyson, H. J.; Wright, P. E. *J. Mol. Biol.* **2004**, *340*, 1131–1142.
(21) Fieber, W.; Kristjansdottir, S.; Poulsen, F. M. *J. Mol. Biol.* **2004**, *339*, 1191–1199.
(22) Meier, S.; Güthe, S.; Kiefhaber, T.; Grzesiek, S. *J. Mol. Biol.* **2004**, *344*, 1051–1069.
(23) Ohnishi, S.; Lee, A. L.; Edgell, M. H.; Shortle, D. *Biochemistry* **2004**, *43*, 4064–4070.
(24) Sallum, C. O.; Martel, D. M.; Fournier, R. S.; Matousek, W. M.; Alexandrescu, A. T. *Biochemistry* **2005**, *44*, 6392–6403.
(25) Ding, K.; Louis, J. M.; Gronenborn, A. M. *J. Mol. Biol.* **2004**, *335*, 1299–1307.
(26) Jensen, M. R.; Markwick, P.; Griesinger, C.; Zweckstetter, M.; Meier, S.; Grzesiek, S.; Bernado, P.; Blackledge, M. *Structure* **2009**, *17*, 1169–1185.
(27) Bernado, P.; Blanchard, L.; Timmins, P.; Marion, D.; Ruigrok, R. W. H.; Blackledge, M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 17002–17007.
(28) Bernado, P.; Bertoncini, C.; Griesinger, C.; Zweckstetter, M.; Blackledge, M. *J. Am. Chem. Soc.* **2005**, *127*, 17968–17969.
(29) Mukrasch, M. D.; Markwick, P. R. L.; Biernat, J.; von Bergen, M.; Bernado, P.; Griesinger, C.; Mandelkow, E.; Zweckstetter, M.; Blackledge, M. *J. Am. Chem. Soc.* **2007**, *129*, 5235–5243.
(30) Dames, S. A.; Aregger, R.; Vajpai, N.; Bernado, P.; Blackledge, M.; Grzesiek, S. *J. Am. Chem. Soc.* **2006**, *128*, 13508–13514.
(31) Jensen, M. R.; Houben, K.; Lescop, E.; Blanchard, L.; Ruigrok, R. W. H.; Blackledge, M. *J. Am. Chem. Soc.* **2008**, *130*, 8055–8061.
(32) Jensen, M. R.; Blackledge, M. *J. Am. Chem. Soc.* **2008**, *130*, 11266–11267.
(33) Meier, S.; Häussinger, D.; Jensen, P.; Rogowski, M.; Grzesiek, S. *J. Am. Chem. Soc.* **2003**, *125*, 44–45.
(34) Meier, S.; Grzesiek, S.; Blackledge, M. *J. Am. Chem. Soc.* **2007**, *129*, 9799–9807.
(35) Kohn, J. E.; Millett, I. S.; Jacob, J.; Zagrovic, B.; Dillon, T. M.; Cingel, N.; Dothager, R. S.; Seifert, S.; Thiyagarajan, P.; Sosnick, T. R.; Hasan, M. Z.; Pande, V. S.; Ruczinski, I.; Doniach, S.; Plaxco, K. W. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 12491–12496.
(36) Merchant, K. A.; Best, R. B.; Louis, J. M.; Gopich, I. V.; Eaton, W. A. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 1528–1533.
(37) Möglich, A.; Joder, K.; Kiefhaber, T. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 12394–12399.
(38) Gabel, F.; Jensen, M. R.; Zaccai, G.; Blackledge, M. *J. Am. Chem. Soc.* **2009**, *131*, 8769–8771.
(39) Wells, M.; Tidow, H.; Rutherford, T. J.; Markwick, P.; Jensen, M. R.; Mylonas, E.; Svergun, D. I.; Blackledge, M.; Fersht, A. R. *Proc. Natl. Acad. Sci. (U.S.A.)* **2008**, *105*, 5762–5767.

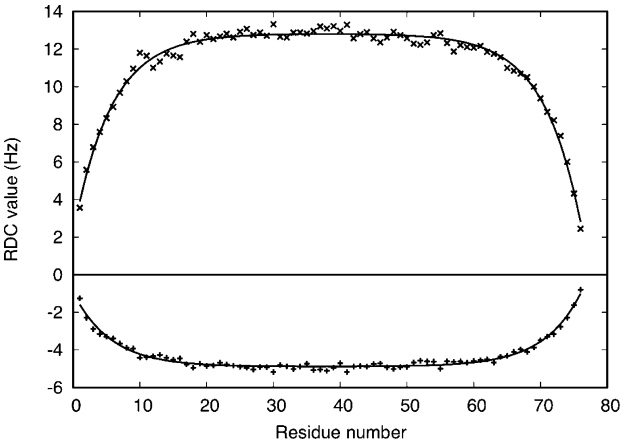

Figure 1. Residual dipolar coupling baselines in unfolded chains. Baseline effects underlying simulated ensemble-averaged RDCs from 100K copies of a polyvaline chain of 76 amino acids in length (crosses) and predicted RDCs following a hyperbolic cosine curve of the form given in eq 1 (line). $^{15}N-^{1}H^{N}$ couplings are shown below zero and $^{13}C^{\alpha}-^{1}H^{\alpha}$ RDCs are shown above zero.

acid-specific or even atomic resolution, as have recently been developed in the Bonvin and Forman-Kay laboratories.[40,41] In order to do this, we develop a novel algorithm to select from a large pool of possible conformers, created using the algorithm flexible-meccano, to best describe the system.

We test this approach, using extensive simulation, to determine how well the fitting of RDCs to reduced conformational ensembles containing few copies of the molecule can correctly reproduce the backbone conformational behavior of the protein. We also use cross-validation of data not employed in the fit to determine the most appropriate ensemble size to characterize the highly fluctuating molecule. Having established approaches that allow accurate mapping of conformational space from RDCs, we apply these methods to the amino acid-specific description of backbone conformational sampling in ubiquitin denatured in 8 M urea at pH 2.5.

Results and Discussion

RDCs from Disordered Proteins Modeled by Multiplication of Local Sampling Profiles and Underlying Baseline. RDCs can be simulated from explicit molecular ensembles of disordered proteins using shape-based considerations of the alignment properties of each copy of the molecule, and the average couplings can be predicted by taking the mean over the entire ensemble.[27,42] Comparison of such predictions with experimental data has revealed the unique sensitivity of RDCs to local and global sampling properties of highly disordered proteins. A key disadvantage of this approach is the number of structures that need to be treated, before the average RDC value converges to a nonfluctuating value. This number can reach many tens of thousands in proteins of 100 amino acids. It has recently been proposed that convergence of RDCs toward experimental data can be achieved with a smaller number of conformers if the protein is divided into short, uncoupled segments (Local Alignment Windows, LAWs) and the RDCs are calculated using the alignment tensor of these segments.[43,44] The ability to describe the conformational properties with ensembles containing fewer structures will of course make any ensemble selection procedure more tractable and is therefore an attractive prospect. In general, however, RDCs are affected both by the local conformational sampling and the chain-like nature of the unfolded protein, which induce an effective baseline reflecting the increasing degrees of freedom available toward the ends of the chain.[45,46] Long-range information is therefore necessarily absent from an approach that only employs LAWs to predict the RDCs. If this approach is employed, the simulated data need to be corrected for the effects of the unfolded chain.

We have simulated ensemble-averaged RDCs for polyvaline chains of differing lengths. The predicted RDCs can be relatively well fitted to a hyperbolic cosine curve of the form (Figure 1)

$$B(i) = 2b\cosh(a(i - d)) - c \quad (1)$$

where i is the residue number and d is half the number of residues. a, b, and c are optimized for each different coupling type, where $(2b - c)$ is the RDC value at position d. This baseline dependence can be used to correct RDCs calculated using LAWs as described below.

RDCs are simulated for the central residue of LAWs of equal length, sliding the LAW one amino acid at a time along the chain (note that the termini are treated in the same way by adding dummy residues beyond the ends of the chain; see Experimental Section). These RDCs are then averaged over all structures. RDCs simulated for LAWs of m amino acids in length will exhibit a flat baseline, because each calculated RDC is at the center of a fragment of m amino acids and is therefore at the middle of the same local effective baseline. The RDC distribution resulting from the LAWs therefore depends on amino acid type but does not contain the baseline effects. It can be shown (Figure 2) that this amino acid-specific distribution can be multiplied with the baseline predicted in eq 1, to closely reproduce RDCs predicted from the explicit full-length description of the protein, which contains both amino acid-specific effects and the chain nature of the full length protein.

In order to determine the convergent characteristics when RDCs are simulated using LAWs of different lengths, we have compared the average values taken over an increasing number of conformers. Examples are shown in Figure 3a of the same $^{1}D_{NH}$ RDC when the RDC is calculated for the central amino acid of LAWs of different lengths (3, 9, 15, 25, and full length protein of 76 amino acids). Further simulations of $^{1}D_{C\alpha H\alpha}$, $^{1}D_{C\alpha C'}$, $D_{NHH\alpha}$, and D_{NHNH} RDCs show similar convergent characteristics (data not shown). It is clear that for the full-length protein the average is only converged when more than 10 000 structures are taken into account, while for LAWs of 15 amino acids this number falls to a few hundred. Figure 3b shows the strong dependence of the range of sampled RDCs on the length of the LAW. As the LAW gets longer, the individual structures can have larger RDC values, rendering the average less and less stable (vide infra).

(40) Marsh, J. A.; Neale, C.; Jack, F. E.; Choy, W.-Y.; Lee, A. Y.; Crowhurst, K. A.; Forman-Kay, J. D. *J. Mol. Biol.* **2007**, *367*, 1494–1510.
(41) Krzeminski, M.; Fuentes, G.; Boelens, R.; Bonvin, A. M. J. J. *Proteins: Struc. Funct. Bioinform.* **2009**, *74*, 894–905.
(42) Jha, A. K.; Colubri, A.; Freed, K.; Sosnick, T. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 13099–13105.
(43) Marsh, J. A.; Baker, J. M. R.; Tollinger, M.; Forman-Kay, J. D. *J. Am. Chem. Soc.* **2008**, *130*, 7804–7805.
(44) Marsh, J. A.; Forman-Kay, J. D. *J. Mol. Biol.* **2009**, *391*, 359–374.
(45) Louhivuori, M.; Pääkkönen, K.; Fredriksson, K.; Permi, P.; Lounila, J.; Annila, A. *J. Am. Chem. Soc.* **2003**, *125*, 15647–15650.
(46) Obolensky, O. I.; Schlepckow, K.; Schwalbe, H.; Solov'yov, A. V. *J. Biomol. NMR* **2007**, *39*, 1–16.

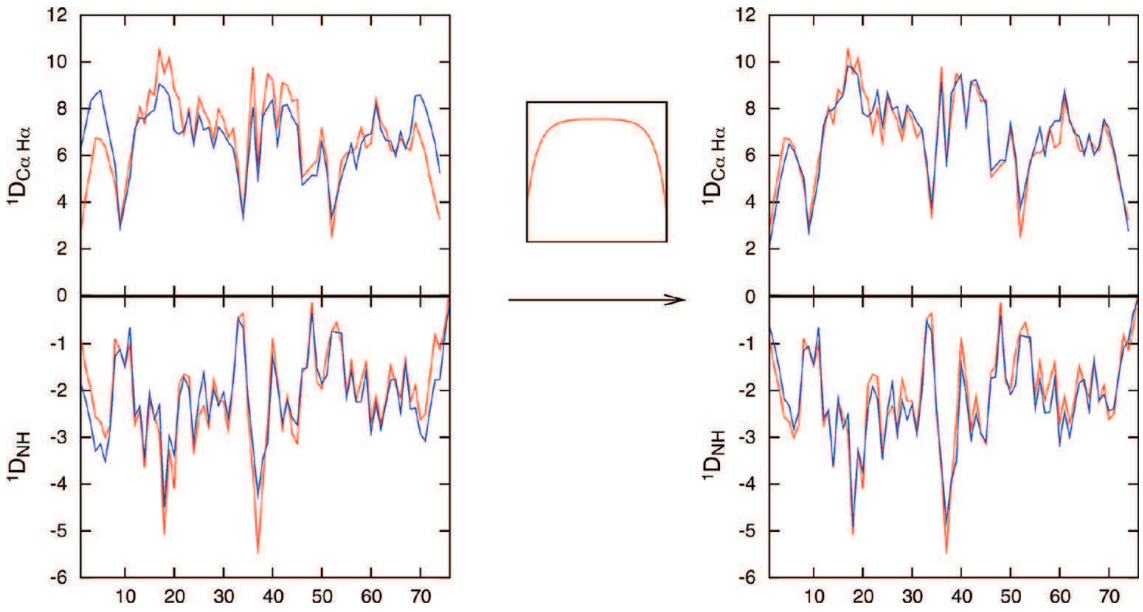

Figure 2. Multiplication of RDCs calculated using LAWs with RDC baselines in unfolded chains. $^{15}N-^1H^N$ and $^{13}C^\alpha-^1H^\alpha$ RDCs calculated from the central amino acid of a 15 amino acid LAW (blue, left) contain no baseline information and therefore diverge from the RDCs calculated from an explicit ensemble using a global alignment tensor (red). When multiplied with the hyperbolic cosine curve (eq 1), RDCs from the LAW (blue, right) more closely resemble the RDCs calculated from the global alignment tensor (red).

Alignment Strand Length Required To Define Accurately Conformational Sampling. In order to further determine the accuracy of describing RDCs using LAWs, we have compared the ability of LAWs of different lengths (after multiplication with the baseline described by eq 1) to reproduce RDCs simulated using a global alignment tensor (Figure 4). Not surprisingly, the shortest LAWs (three amino acids in length) never correctly reproduce average RDCs, due to the effects of neighboring amino acids (beyond nearest neighbors), on the local conformational sampling. The influence of neighboring residues on local conformational sampling is commonly estimated in terms of a so-called "persistence length", beyond which the remainder of the chain can be considered to exert a negligible effect. The persistence length depends on the relative rigidity of the local primary sequence. The relevance of taking full account of the persistence length on the local conformational sampling is further demonstrated by simulations that have been performed using a more rigid statistical coil model for which RDCs simulated using LAWs of nine amino acids fail to reproduce the averaged RDCs calculated using the global alignment tensor (data not shown). These simulations therefore indicate that while convergence characteristics of the predicted RDCs improve with shorter LAWs, the shortest strands can never fully reproduce the correct average, even if a very large number of structures were used in the average. On the basis of these simulations, we consider that a LAW length of 15 amino acids should be an acceptable compromise between efficiency and accuracy for the subsequent analyses.

How Many Structures Are Required for RDCs To Define Accurately Conformational Sampling? The next question concerns the number of structures required to describe correctly the conformational sampling. The averaging of RDCs is particularly demanding in terms of numbers of structures for two main reasons: first, because of the large number of backbone dihedrals whose relevant conformational space must be efficiently sampled before the overall shape and dimensions of the protein, and therefore the associated alignment tensor, average to convergent values. A second consideration is less obvious, but potentially more important: each dipolar coupling calculated from a single conformer of the entire molecule will sample a value within a range that can be orders of magnitude higher than the range spanned by the average values (Figure 3b). This dynamic-range problem can induce significant instability in the fitting procedure when using an ensemble containing too few structural models.

In order to numerically estimate the minimum number of structures that can accurately reproduce the true structural propensities of a conformational equilibrium, we have undertaken the following simulation: Two distinct statistical coil sampling regimes were defined, and entire sets of RDCs were calculated from flexible-meccano using these regimes with the global alignment tensor. The first, regime S, defines the standard statistical coil model employed in flexible-meccano, where amino acid-specific conformational distributions are extracted from populations of coil regions found in the protein structural database. The second sampling regime (E) samples a more extended region of Ramachandran space, populating the region $\{50° < \psi < 180°\}$ with a higher propensity than the S regime (see Experimental Section), while retaining the amino acid specific sampling from the S database. These data sets were then used as targets for the ensemble selection algorithm ASTEROIDS (A Selection Tool for Ensemble Representations Of Intrinsically Disordered States) described in the Experimental Section.

The ability of the algorithm to reproduce the correct conformational sampling and the correct RDCs for two different LAWs and the global alignment tensor is summarized in Figure 5 as a function of the number of structures constituting the ensemble. Using the target function χ^2_{Ram}, which measures the population of four different regions of Ramachandran space defined in Figure 6, we measure the ability of the protocol to reproduce amino acid-specific conformational sampling throughout the molecule (see Experimental Section). In each of the three considered window lengths, (9, 15, and full length protein), the

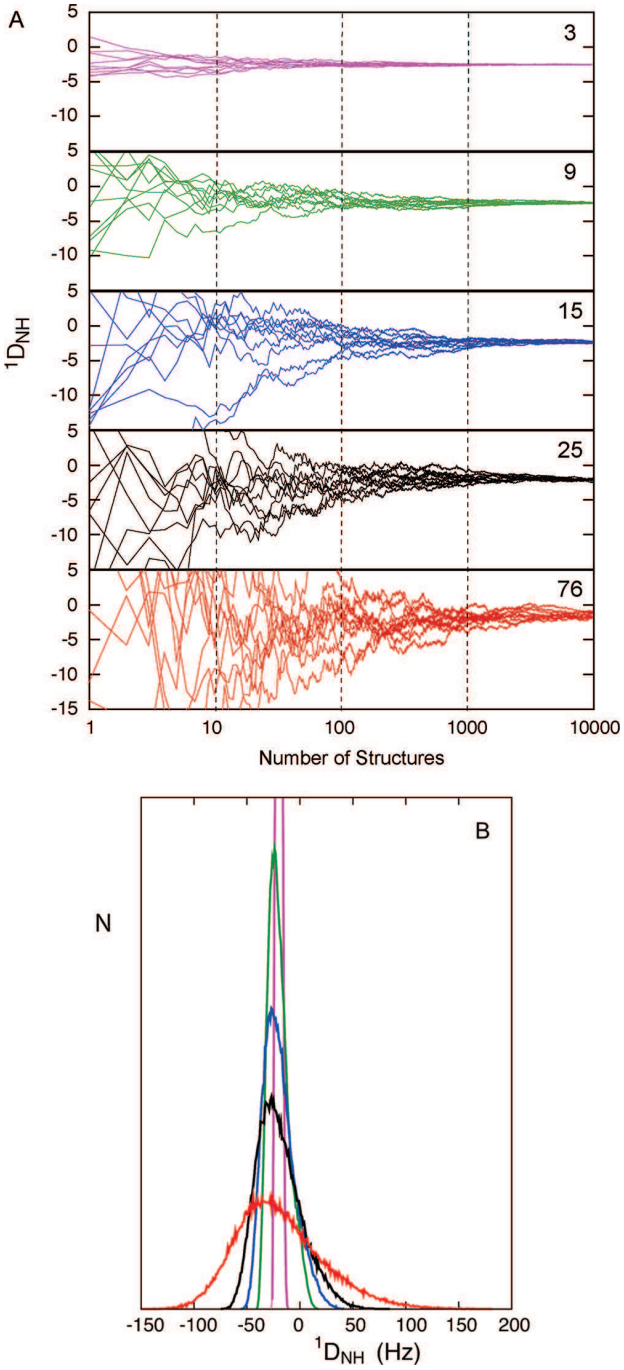

Figure 3. Convergence of $^{15}N-^1H^N$ RDCs calculated using LAWs of different lengths. (a) Comparison of 10 simulations of the central amino acid of an m amino acid LAW. The same $^1D_{NH}$ RDC (amino acid 41 of ubiquitin) is calculated using LAWs of $m = 3$, 9, 15, 25 or from the full length (76 amino acid) protein using a global alignment tensor. The x-axis represents the number of structures used to calculate the average. (b) Range and distribution of RDCs from the simulations shown in part a. Color code is the same in both cases (purple, three amino acid window; green, nine amino acids; blue, 15 amino acids; black, 25 amino acids; red, 76 amino acids).

reproduction of the RDCs improves rapidly with the number of structures included in the ensemble average. Simultaneously, the reproduction of the correct conformational sampling (the sampling used to simulate the RDC data) improves in all cases. These simulations, and those applied to the more extended

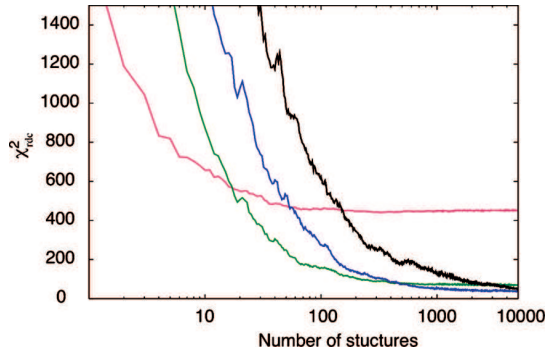

Figure 4. Accuracy of RDCs calculated using LAWs compared to a full length description. Equation 4 was used to directly compare the ability of RDCs calculated using the convolution of baseline and LAWs to reproduce RDCs calculated using an explicit description of the full length protein. The x-axis defines the number of averaged RDCs. χ^2_{RDC} was calculated over the entire protein. Color code: purple, three amino acid window; green, nine amino acids; blue, 15 amino acids; black, 25 amino acids.

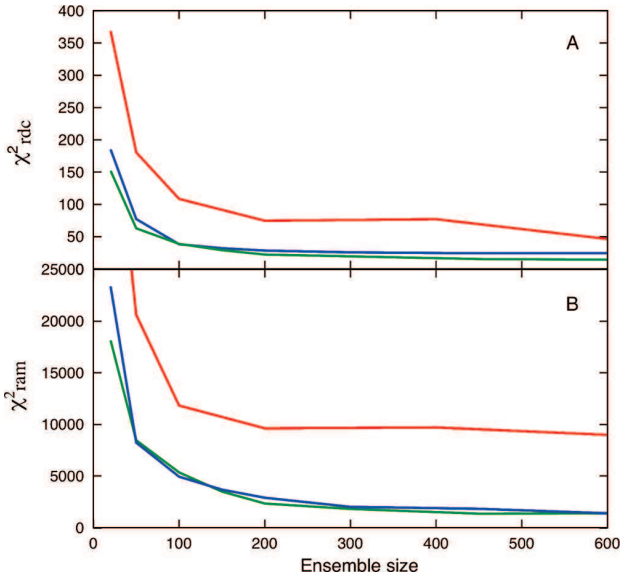

Figure 5. Accuracy of ensembles of structures calculated using LAWs of different lengths. The ability of ASTEROIDS to reproduce the correct conformational sampling and the correct RDCs for LAWs of different lengths is summarized as a function of the number of structures constituting the ensemble. (a) χ^2_{RDC} measures the reproduction of the target RDCs calculated using the full length 50 000-strong explicit description of the global alignment tensor. (b) χ^2_{Ram} measures the ability of the protocol to reproduce conformational sampling throughout the molecule. Color code: green, nine amino acid LAWs; blue, 15 amino acid LAWs; red, 76 amino acids (global alignment tensor). The x-axis defines the number of structures used.

sampling regime (data not shown), indicate that the optimal combination for an accurate description of conformational behavior of the protein backbone requires a window length of at least 15 amino acids and 200 structures.

The site-specific reproduction of the different RDCs comprising the χ^2_{RDC} using an ensemble of 200 and 20 structures is shown in Figure 7, for a LAW of 15 amino acids. Although the fit is significantly poorer in the case of 20 structures, the overall features are actually quite well reproduced, and the quality of the fit would probably be considered acceptable in the presence of commonly encountered levels of experimental noise. The conformational sampling is, however, very poorly reproduced, throughout the protein, when only 20 structures are

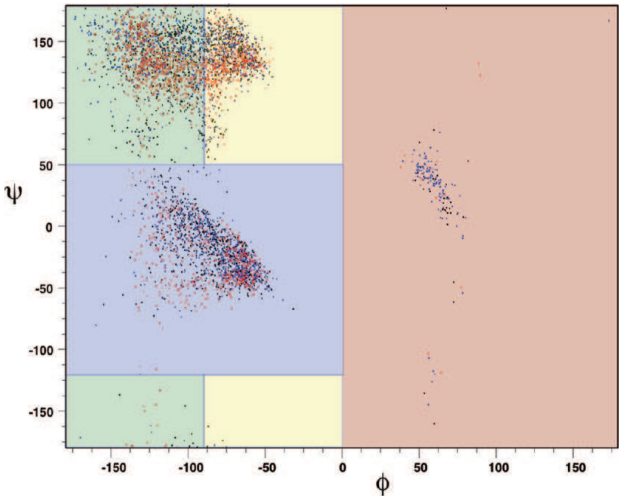

Figure 6. In order to quantify the similarity between conformational sampling between different ensembles, Ramachandran space is divided into four quadrants and defined as follows: α_L, $\{\phi > 0°\}$; α_R, $\{\phi < 0, -120° < \psi < 50°\}$; β_P, $\{-90° < \phi < 0°, \psi > 50°$ or $\psi < -120°\}$; β_S, $\{-180° < \phi < -90°, \psi > 50°$ or $\psi < -120°\}$. The population of these quadrants is indicated as p_{α_L}, p_{α_R}, p_{β_P}, and p_{β_S}. Dots represent standard statistical coil distributions of valine (red), lysine (blue), and leucine (black).

included. This is graphically underlined in Figure 8, where the populations of the four quadrants of conformational space present in the 200- and 20-fold ensembles are compared with those present in the ensemble used to create the simulated data. Discrepancies in the population of the different quadrants of up to 30% compared to the value present in the original ensemble are found throughout the primary sequence for the 20-fold ensemble. These differences do not appear to be correlated to amino acid type. The 200-fold ensembles, on the other hand, closely reproduce the original sampling (figure 8b) for every region of primary sequence. It is therefore evident that, in cases where too few structures are included in the average, achieving acceptable reproduction of experimental data does not guarantee that the resulting ensemble accurately represents the correct conformational distribution.

Application of ASTEROIDS to Experimental RDCs from Urea-Unfolded Ubiquitin. Using the optimal parameters determined on the basis of the simulations described above, we have applied the ASTEROIDS approach to the determination of a representative ensemble to describe the conformational behavior of the protein ubiquitin under denaturing conditions (pH 2.5 and 8 M urea). In the initial analysis, ensembles of 200 structures were selected from a set of 12 000 conformers for which LAWs of 15 amino acids in length were used to calculate the dipolar couplings. The results, shown in Figure 9a, indicate a reasonable reproduction of experimental data but reveal notable systematic effects, in particular that the $D_{\mathrm{NHH}\alpha(i-1)}$, $D_{\mathrm{NHNH}(i+1)}$ RDCs are overestimated when the other couplings, effectively the $^1D_{\mathrm{NH}}$ and $^1D_{\mathrm{C}\alpha\mathrm{H}\alpha}$ RDCs agree optimally with simulation. These observations agree qualitatively with identification of differential scaling of $^1\mathrm{H}-^1\mathrm{H}$ couplings compared to covalently bound spins in the analysis of these RDCs. In order to allow for this possibility in the current analysis, we allowed for two independent scaling factors, K_1 for the $^1D_{\mathrm{NH}}$, $^1D_{\mathrm{C}\alpha\mathrm{H}\alpha}$, and $^1D_{\mathrm{C}\alpha\mathrm{C}'}$ and K_2 for the $D_{\mathrm{NHH}\alpha}$, $D_{\mathrm{NHH}\alpha(i-1)}$, $D_{\mathrm{NHNH}(i+1)}$, and $D_{\mathrm{NHNH}(i+2)}$. These factors are optimized uniformly for the covalently bound and through-space dipolar interactions, resulting in the data reproduction shown in Figure 9b. The two scaling factors $K_1 = 0.58$ and $K_2 = 0.96$ differ by approximately 0.6, a difference that may result from additional local conformational dynamics that are not taken into account by the statistical coil model and that scale the $D_{\mathrm{NHH}\alpha(i-1)}$, $D_{\mathrm{NHNH}(i+1)}$ RDCs differentially to the RDCs between spins whose distances are effectively fixed. This possibility is currently under more detailed investigation.

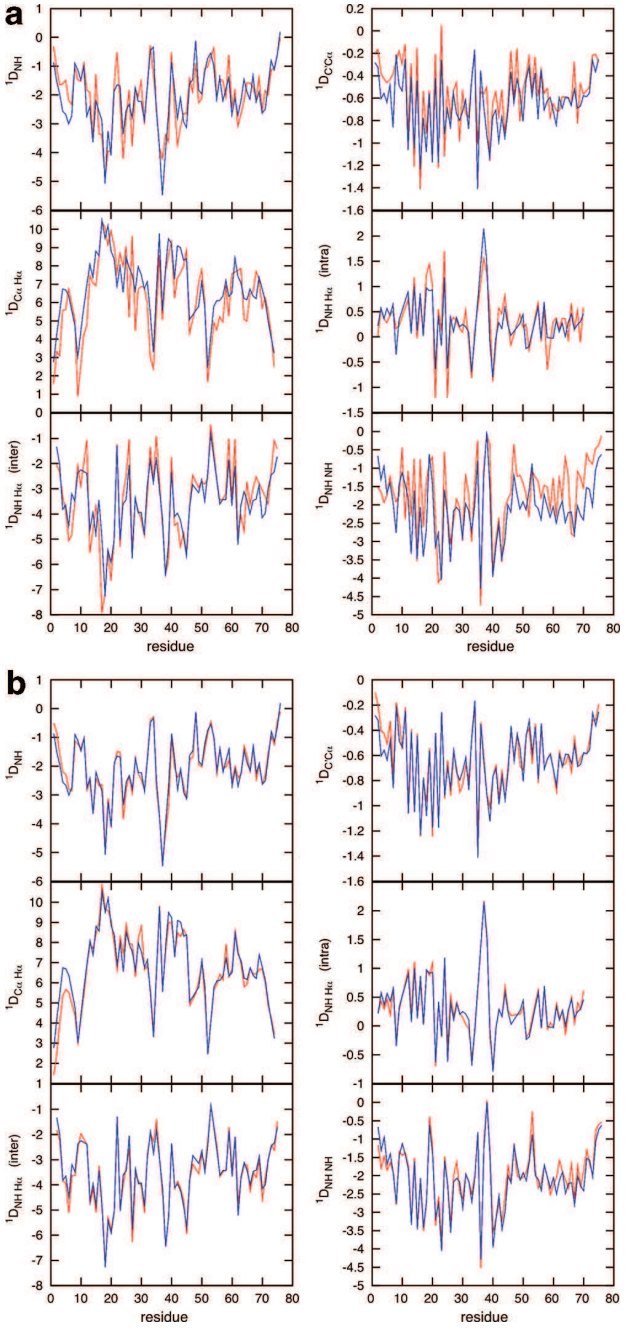

Figure 7. Site-specific reproduction of the RDCs simulated using an explicit ensemble of 50 000 structures. (a) Reproduction of the target data (blue) using an ensemble of 20 structures (red) for a window length of 15 amino acids. (b) Reproduction of the target data (blue) using an ensemble of 200 structures (red) for a window length of 15 amino acids. In both cases, the genetic algorithm ASTEROIDS was used to select the optimal ensemble.

In order to test the validity of the approaches shown here for the analysis of experimental data, we have repeated the ASTEROIDS ensemble selection procedure, taking 10% of the

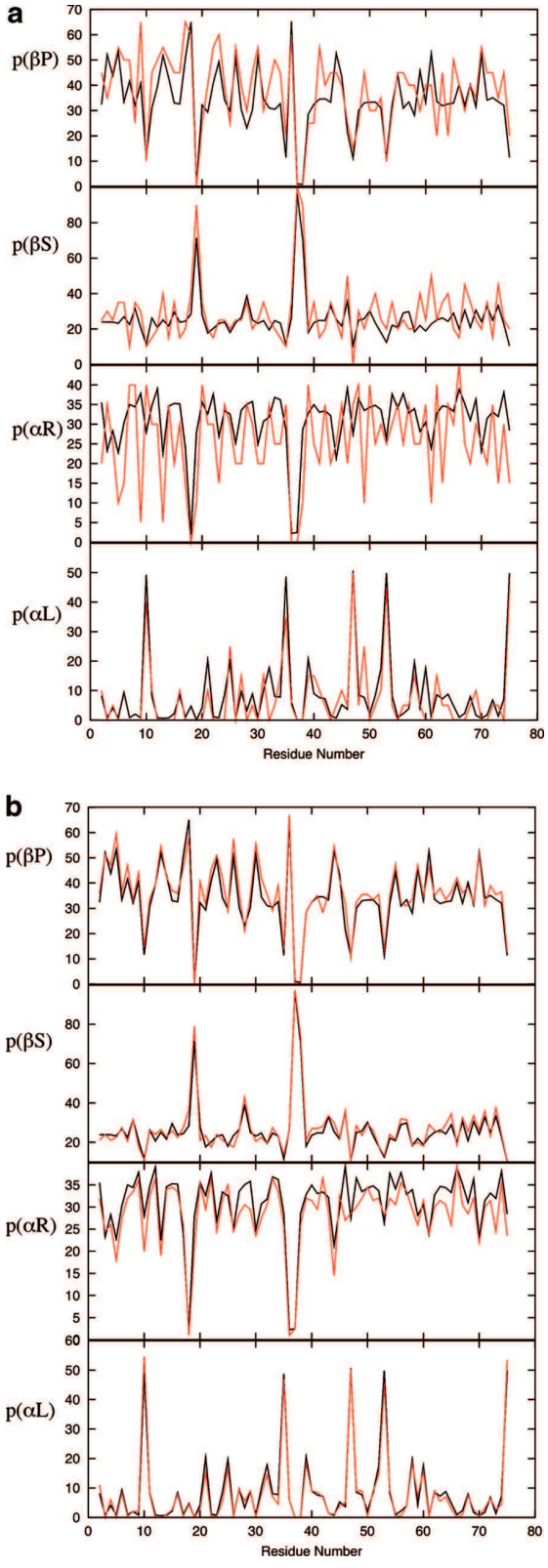

Figure 8. Accuracy of the reproduction of conformational sampling using the ASTEROIDS approach with ensembles of 20 and 200 structures. Populations of the four quadrants of conformational space defined in Figure 6 using the (a) 20-fold and (b) 200-fold ensembles (red) compared with those present in the ensemble used to create the simulated data (black). Discrepancies in the population of the different quadrants of up to 30% compared to the value present in the original ensemble are found for ensembles of size 20.

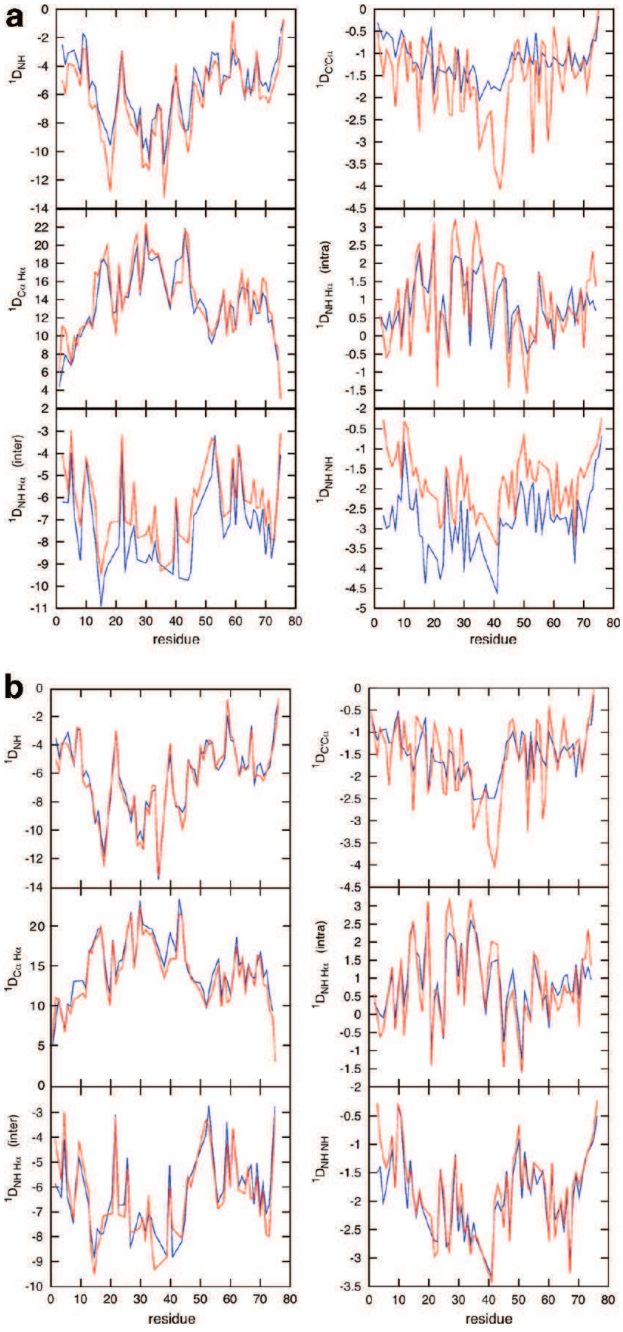

Figure 9. Application of ASTEROIDS to experimental RDCs from urea-unfolded ubiquitin. (a) Reproduction of experimental data (red) using an ensemble of 200 structures (blue). (b) Reproduction of experimental data (red) using an ensemble of 200 structures (blue) with differential scaling of the covalently bound and interproton RDCs.

RDCs out of the analysis and comparing the predicted values using the resulting ensemble with the experimental RDCs. The results are shown in Figure 10, where the back-calculated RDCs are found to be in reasonable agreement with the experimentally determined values. The calculation was repeated 10 times at seven different ensemble sizes. The average cross-validated χ^2 is plotted as a function of ensemble size (Figure 10b). The size of 200-fold ensembles used in the current approach is within the range where the cross validation target function is essentially flat.

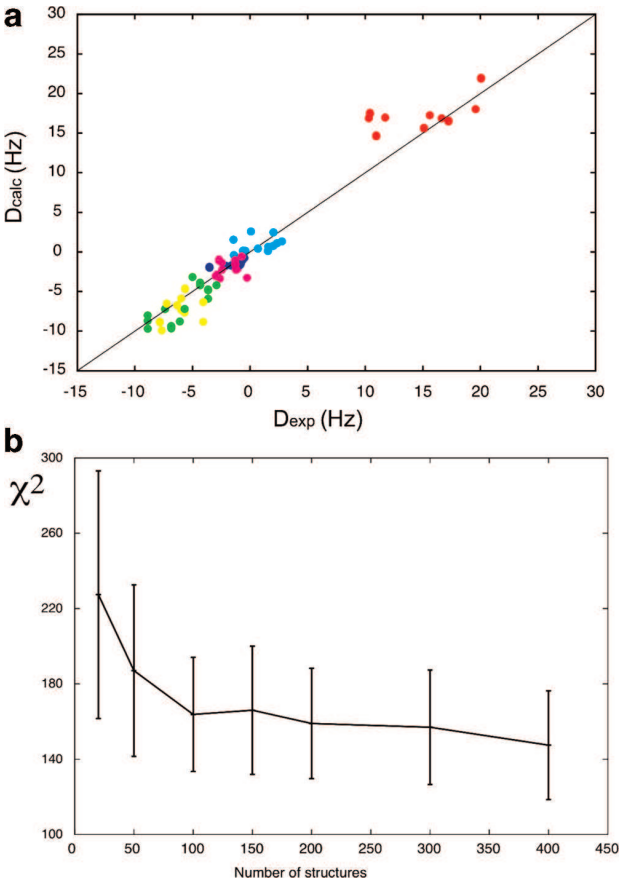

Figure 10. Reproduction of data not used in the fitting procedure. (a) The ASTEROIDS ensemble selection procedure was repeated, taking 10% of the RDCs out of the analysis and comparing the predicted values using the resulting ensemble with the experimental RDCs. Color code: green, $^1D_{NH}$; red, $^1D_{C\alpha H\alpha}$; dark blue, $^1D_{C\alpha C'}$; cyan, $D_{NHH\alpha}$; yellow, $D_{NHH\alpha(i-1)}$; magenta, $D_{NHNH(i+1)}$.(b) Average χ^2 over 10 cross-validation calculations at each of seven different ensemble sizes. The 200-fold ensemble size used in the current approach is within the range where the cross-validation target function is essentially flat.

The precision with which the RDCs can define the conformational behavior of the backbone has been assessed using noise-based Monte Carlo simulations (see Experimental Section) based on estimates of experimental uncertainty. The results are summarized in Figure S1 of the Supporting Information, and show that the average uncertainty in the populations of the different quadrants is approximately ±3%. We have also repeated the entire analysis in the absence of one experimental data set to assess the relative importance of each data set for the conformational description. The results are shown in Figure S2 and summarized in Tables S1 and S2 of the Supporting Information, where the backbone sampling is compared to the populations determined using all data. The root-mean-square deviation of the four populations defined in Figure 6 and the average differences demonstrate that although we find that the most important RDCs are the $D_{NHHai,i+1}$ and D_{NHNH}, the effects are actually not very large when these RDCs are removed (maximum rmsd of 5%, and average difference in populations of 3%). These results suggest that both covalently bound and interproton RDCs are important for an accurate description of conformational sampling but that none of the RDC types are critical for the validity of the description or the conclusions drawn from it.

The amino acid Ramachandran sampling has been used to calculate expected $^3J_{NHH\alpha}$ scalar couplings, reporting on the sampling of the ϕ backbone dihedral angle. These values have been compared to experimentally determined couplings[47] (Figure S3, Supporting Information), in comparison to the reproduction of the data using the standard coil database. The *J*-coupling data reproduction is quite good in both cases, but only slightly better in the case of the selected ensemble ($\chi^2 = 11.5$ compared to 12.6), probably reflecting the fact that the differences in the two descriptions are often found in the distribution of the ψ backbone dihedral angle. However, this analysis does demonstrate that the local analysis of RDCs in terms of Ramachandran distributions does not contradict independent experimental data in a significant way.

Urea Preferentially Affects the Conformational Sampling of Amino Acids with Side Chain Hydrogen-Bonding Moieties. Figure 11 shows the backbone dihedral angle distributions resulting from the analysis of experimental data of urea-unfolded ubiquitin and the normalized difference compared to the distribution of angles derived using an ensemble of structures produced using the standard statistical coil model of the unfolded state. Figure S4 of the Supporting Information shows the amino acid specific populations of all amino acids for the standard statistical coil model. The sampling of the different regions of the Ramachandran space defined in Figure 6 is summarized in Figure 12.

In general, the results indicate that the sampling of backbone dihedral angles in Ramachandran space is more extended, sampling the β_P and β_S regions with higher propensity and the α_R region with lower propensity than the statistical coil database. This result is in agreement with a previous study of the more general characteristics of conformational sampling, using the same experimental data.[34] In this study, a hypothesis-driven approach was used to suggest a general extension of conformational sampling of the peptide chain. With the new techniques developed here, we are able to extract amino acid-specific conformational sampling directly from the RDC data. This approach relies on the supposition that the database from which structures are selected contains enough conformational diversity to allow for a representative description to be constructed from its population. Under these conditions, the method is relatively hypothesis-free in comparison to previous approaches. This reveals that the effects of urea on backbone conformational sampling are far from uniform. The extended nature of the chain is more apparent in localized contiguous segments of primary sequence: the regions 30−36 and 70−73 sample the β_P region more extensively than both the statistical coil and the remainder of the protein, while extended β regions are preferentially sampled in the region 14−18. This latter tendency may be correlated with the previously observed presence of a small (around 20%) residual population of β hairpin in this region of the molecule.[48] Amino acids preceding prolines (18 and 36) are found to better reproduce experimental RDCs with a more uniform sampling of propensities in the β_P and β_S regions, compared to the statistical coil database that preferentially samples the α_R region.

The comparison with the statistical coil model clarifies detail that may be masked by amino acid-specific sampling of backbone dihedral angle and allows the identification of sites

(47) Peti, W.; Henning, M.; Smith, L. J.; Schwalbe, H. *J. Am. Chem. Soc.* **2000**, *122*, 12017–12018.
(48) Meier, S.; Strohmeier, M.; Blackledge, M.; Grzesiek, S. *J. Am. Chem. Soc.* **2007**, *129*, 754–755.

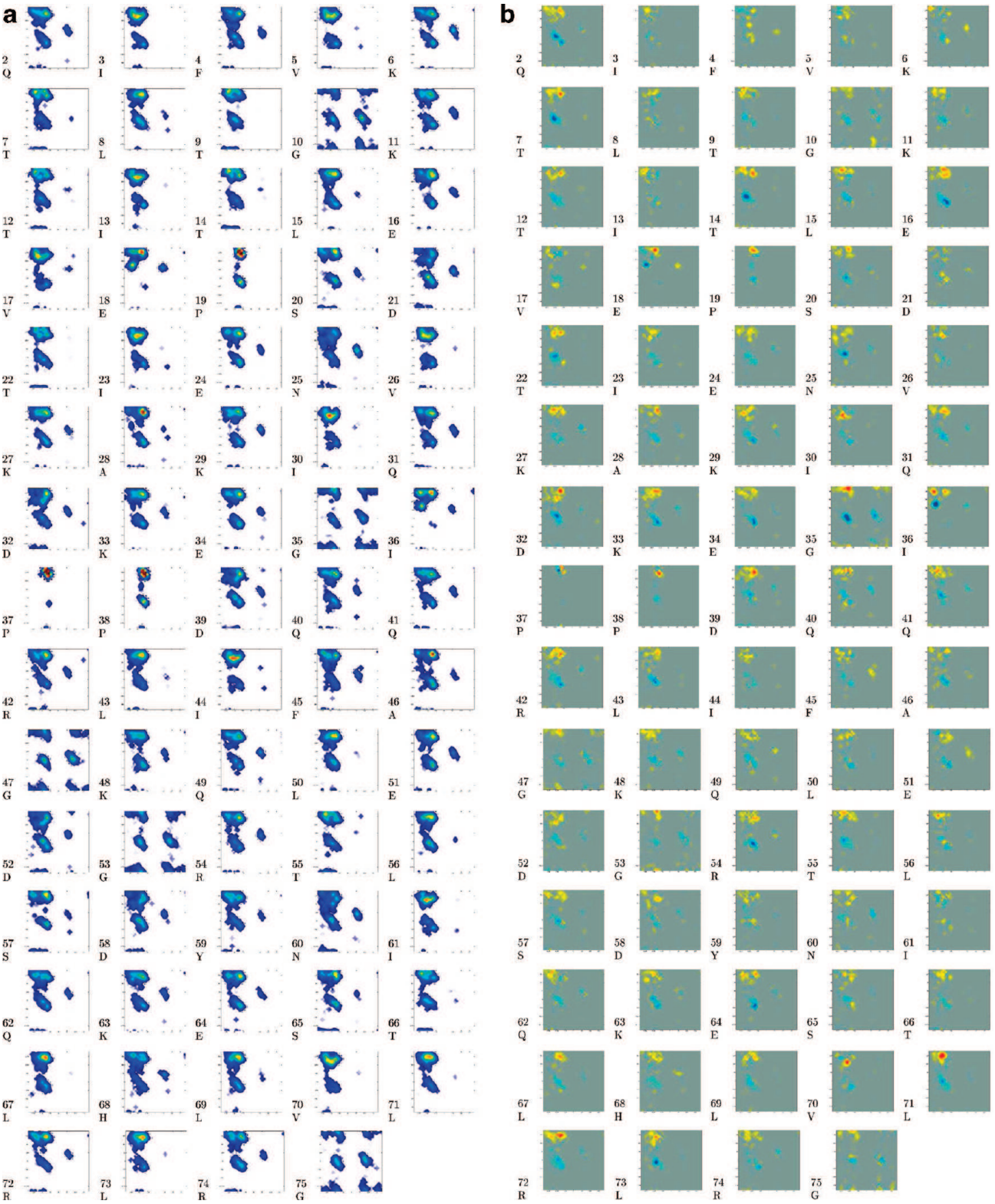

Figure 11. Amino acid-specific Ramachandran distributions for unfolded ubiquitin in 8 M urea at pH 2.5 in comparison with a standard statistical coil distribution. The populations increase from dark blue, via cyan, green, and yellow, to red. (a) Conformational sampling determined from the ASTEROIDS analysis of experimental RDC data (10 calculations were combined to produce 2000 conformers for the sake of figure resolution). (b) Difference between the conformational sampling distributions shown in panel a and the conformational sampling for the flexible-meccano statistical coil distribution. In this case, blue to green corresponds to negative values (population is lower in the urea unfolded sampling than in the statistical coil) and green (via yellow) to red corresponds to positive values (population is higher in the urea-unfolded sampling than in the statistical coil). Gray corresponds to equal populations.

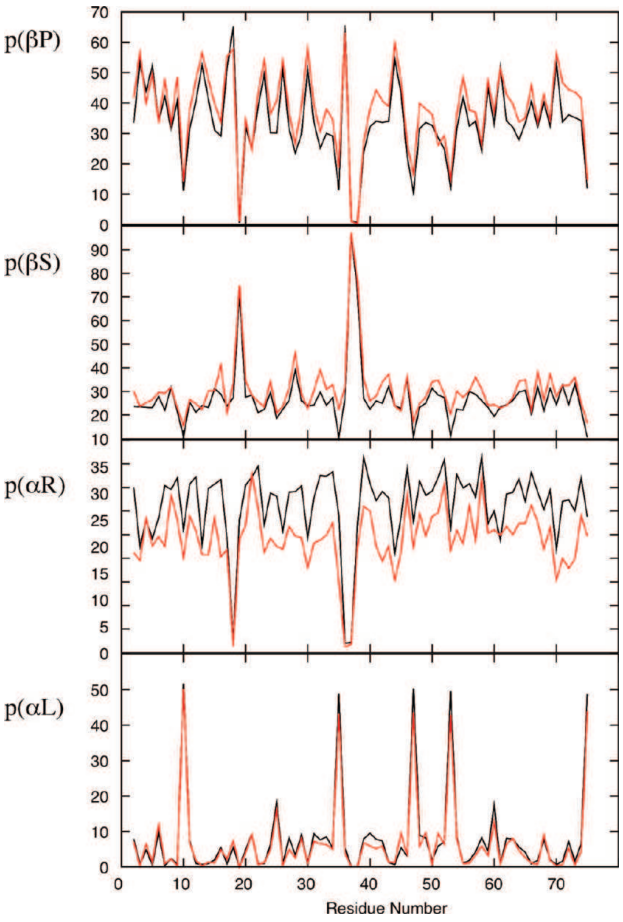

Figure 12. Populations of the four quadrants of conformational space defined in Figure 6 using the amino acid-specific Ramachandran distributions for unfolded ubiquitin in 8 M urea at pH 2.5 shown in Figure 11 (red) in comparison to a standard statistical coil distribution (black).

whose behavior deviates from random coil in the presence of urea. In this context, it is interesting to note that the amino acids whose backbone conformational sampling are most systematically affected by the presence of urea are threonine (four out of seven have a notably more extended backbone sampling than in the statistical coil model), glutamic acid (three out of five are more extended than the statistical coil model), and arginine (three out of four are more extended than the statistical coil model). These amino acids all contain potential hydrogen-bond-donor moieties on their side chains. A recent study using vibrational spectroscopy demonstrated that at low pH urea orients with the carboxyl group pointing toward the protein surface, an observation that supports the suggestion that hydrogen-bond-donor groups may interact preferentially with urea.[49] By contrast, only three of a total of 24 hydrophobic amino acids (valine, leucine, isoleucine, alanine, tyrosine, and phenylalanine) exhibit significantly different conformational sampling between the urea-denatured and the statistical coil states. The specific amino acid composition may therefore be responsible for the apparent localization of differential backbone sampling properties in the different regions of the protein. A recent study used small angle scattering to estimate the number of additional urea molecules that are preferentially recruited during the unfolding transition of ubiquitin from neutral to acidic pH to be approximately 20, a number that correlates qualitatively with the observation here that the backbone behavior of approximately a third of the amino acids are preferentially affected by the presence of urea.[38]

Conclusions

In this study, we have used extensive simulation to optimize an approach that exploits experimental RDCs measured from unfolded proteins to determine conformational sampling on an amino acid-specific basis. Previous applications have used a full-length description of the protein, averaging RDCs over an unrestrained ensemble that is large enough to allow for convergence of the coupling values. Although providing important insight into the behavior of a number of disordered proteins for which conformational information is otherwise difficult to measure, these studies are hypothesis-based, testing different conformational sampling regimes and comparing them to experimental data, an approach that severely limits both the scope and application as well as the potential for discovery. Here we develop a general approach that allows one to select an ensemble directly from the experimental data. Our combination of analytical baseline descriptor and numerical averaging of smaller alignment windows is tested against simulation, and on the basis of these simulations, parameters such as window length and number of structures are calibrated. We find that a combination of LAWs of 15 amino acids in length, with ensemble sizes of 200, accurately describes conformational space, while ensembles of 20 structures reproduce the experimental data but, critically, do not reproduce the correct conformational sampling. Using this approach we can describe conformational sampling at an amino acid resolution.

These approaches have been applied to the amino acid-specific description of backbone conformational sampling in ubiquitin denatured in 8 M urea at pH 2.5. Having established the precision that the approach is expected to offer, we are able to analyze in fine detail the local conformational differences between the standard statistical coil description and the sampling defined by the experimental data measured in the presence of urea, and we interpret this in the context of urea binding or interacting with specific types of amino acids in the peptide chain.

Experimental Section

Experimental methods for measuring the RDCs included in the analysis have been presented elsewhere. All data were taken from the earlier study by Meier et al.[34]

Flexible-Meccano Calculations. Simulated RDCs were calculated using the program flexible-meccano interfaced to the program PALES[50] as described. The program was run in two modes: For calculations using a global alignment tensor for the entire molecule, the standard procedure was used. For calculations using the local alignment windows (LAWs) the RDC for the central amino acid of the local *m* amino acid segment (3, 9, 15, or 25) was calculated for each individual structure. For the terminal amino acids, alanine amino acids were added to the N or C terminus during the building of the protein, such that the *m* amino acid segment was always present. The resulting RDC profile along the primary sequence is calculated by averaging each value and multiplying with the effective baseline given in eq 1. If RDCs were calculated using the full length protein, they were averaged over all conformers as previously described.

(49) Chen, X.; Sagle, L. B.; Cremer, P. S. *J. Am. Chem. Soc.* **2007**, *129*, 15104–15105.

(50) Zweckstetter, M.; Bax, A. *J. Am. Chem. Soc.* **2000**, *122*, 3791–3792.

A pool of 12 000 structures is generated with flexible-meccano. Half of the structures were calculated using the standard statistical coil model S, and the other half using a more extended regime E. The sampling regime (E) samples a more extended region of Ramachandran space, populating the region $\{50° < \psi < 180°\}$ with a higher propensity than the S regime (78% compared to 59%).

ASTEROIDS Ensemble Selection. ASTEROIDS uses a genetic algorithm[51−53] to build a representative ensemble of structures of fixed size N from a large database. The algorithm selects an ensemble of N structures using the following fitness function compared to the experimental data.

$$\chi^2_{\text{asteroids}} = \sum_i w_i^2 (D_{i,\text{calc}} - D_{i,\text{exp}})^2 \quad (2)$$

where w_i is the weight of coupling D_i. The weights were set according to coupling type and determined by the range of each type of coupling in hertz. Values of w were set to 1.0 for $^1D_{\text{NH}}$ and $D_{\text{NHH}\alpha(i-1)}$, 0.5 for $^1D_{\text{C}\alpha\text{H}\alpha}$, 2.0 for $^1D_{\text{C}\alpha\text{C}'}$, $D_{\text{NHH}\alpha}$, and $D_{\text{NHNH}(i+1)}$, and 3.0 for $D_{\text{NHNH}(i+2)}$. The final ensemble is obtained from generations of ensembles that undergo evolution and selection using this fitness function. Each generation comprises 100 different ensembles of size N.

Evolution can proceed in three different ways: random, mutation, and crossing. At each evolution step, the protocol ensures that a structure does not appear more than once in a given ensemble and that a given ensemble is not repeated in a generation. Random evolution proceeds by randomly selecting structures in the complete database. Mutation occurs by taking an ensemble and replacing 1% of the structures (or at least one structure) by structures randomly selected from the complete database (external mutation) or from a new database containing all the structures selected at least once in the previous generation (internal mutation). Crossing is achieved by randomly pairing ensembles from the previous generation. New ensembles are generated by selecting N structures in a pool made of the structures present in the previously defined pairs.

The first generation is always obtained using random evolution. Evolution of this generation is achieved by the following procedure. New ensembles are generated (100 by random evolution, 100 by external mutation, 100 by internal mutation and 100 by crossing). Among these new ensembles and the previous generation, 100 different ensembles representing minima with respect to the fitness function are selected using tournaments to provide the next generation. Ensembles are randomly split into groups and then ordered using the fitness function to determine the winners of the tournament. The best ensembles of each tournament are retained to form the next generation. The number of tournaments and the number of winners of each tournament are adjusted such that 100 ensembles are selected. Selection pressure increases as the number of tournaments decreases. To avoid premature convergence in local minima, the selection pressure is gradually increased during evolution. The number of tournaments therefore successively goes from 100 to 50, 25, 20, 10, 2, and to 1. To ensure robustness of the fitting procedure, the evolution and selection processes are repeated over 2000 successive generations.

Ramachandran Segment Division. In order to describe the sampling of conformational space in the different ensembles and their agreement with known distributions, Ramachandran space is divided into four quadrants indicated in Figure 6 and defined as follows: α_L, $\{\phi > 0°\}$; α_R, $\{\phi < 0, -120° < \psi < 50°\}$; β_P, $\{-90° < \phi < 0°, \psi > 50°$ or $\psi < -120°\}$; β_S, $\{-180° < \phi < -90°, \psi > 50°$ or $\psi < -120°\}$.

The population of these quadrants is indicated as p_{α_L}, p_{α_R}, p_{β_P}, and p_{β_S}. The Ramachandran similarity factor χ^2_{Ram} of the entire molecule is measured by the following function:

$$\chi^2_{\text{Ram}} = \sum_i \sum_q (p_{i,q,\text{ref}} - p_{i,q,\text{fit}})^2 \quad (3)$$

where p_q are the four different populations of the quadrants q, i are the different amino acids, and $_{\text{ref}}$ and $_{\text{fit}}$ signify the target and fitted Ramachandran distributions.

Comparison of RDCs. In order to compare RDCs calculated using different window lengths with those calculated using 50 000 conformers from the full length description of the protein, the following function χ^2_{RDC} is used:

$$\chi^2_{\text{RDC}} = \sum_i (D_{i,\text{LAW}} - D_{i,\text{fl}})^2 \quad (4)$$

where $D_{i,\text{LAW}}$ represents the RDC calculated using LAWs, after multiplication with the baseline function given in eq 1, and $D_{i,\text{fl}}$ is the RDC calculated using the full length description.

Monte Carlo Simulations and Error Analysis. In order to estimate the precision with which the conformational sampling can be defined on the basis of experimental RDCs, we have run noise-based Monte Carlo simulations, using random sampling of Gaussian distributions whose width is based on experimentally estimated uncertainties for each RDC. Fifty Monte Carlo simulations were run, and the effective uncertainty of the Ramachandran quadrant population was calculated on the basis of this.

In order to estimate the importance of the different RDC types, we have repeated the analysis of experimental data with one entire data set removed from the ASTEROIDS approach.

***J*-Coupling Analysis.** $^3J_{\text{NHH}\alpha}$ scalar couplings were calculated by averaging over the amino acid-specific ϕ backbone dihedral angle distributions and compared to experimentally measured values, using recently derived Karplus relationships.[54]

Acknowledgment. L.S. received a grant from the French Ministry of Education. This work was supported by the French Research Ministry through ANR-PCV07_194985. M.R.J. benefited from an EMBO fellowship and Lundbeckfonden support.

Supporting Information Available: A figure showing the standard statistical coil distribution on a residue-specific basis. Residue-specific populations of Ramachandran space resulting from Monte Carlo simulations. A figure and tables showing conformational sampling of the different quadrants of Ramachandran space when specific RDC types are removed. A figure showing calculated and experimental ^3J scalar couplings. This material is available free of charge via the Internet at http://pubs.acs.org.

JA9069024

(51) Fraser, A. S. *Austr. J. Biol. Sci.* **1957**, *10*, 484–491.
(52) Holland, J. H. *Adaptation in Natural and Artificial Systems*; University of Michigan Press: Ann Arbor, 1975.
(53) Jones, G. *Genetic and Evolutionary Algorithms. Encyclopedia of Computational Chemistry*; Wiley: Chichester, U.K., 1998.
(54) Markwick, P. R. L.; Showalter, S. A.; Bouvignies, G.; Brüschweiler, R.; Blackledge, M. *J Biomol NMR* **2009**, *45*, 17–21.

J|A|C|S COMMUNICATIONS

Published on Web 01/11/2010

Defining Conformational Ensembles of Intrinsically Disordered and Partially Folded Proteins Directly from Chemical Shifts

Malene Ringkjøbing Jensen,* Loïc Salmon, Gabrielle Nodet, and Martin Blackledge*

Protein Dynamics and Flexibility, Institut de Biologie Structurale Jean-Pierre Ebel, CEA, CNRS, UJF, UMR 5075, 41 Rue Jules Horowitz, Grenoble 38027, France

Received November 25, 2009; E-mail: malene.ringkjobing-jensen@ibs.fr; martin.blackledge@ibs.fr

A significant fraction of proteins - over 40% of the human proteome - are not folded, or are only partially folded in their functional form.[1] These intrinsically disordered proteins (IDPs) are strongly implicated in important human pathologies such as cancer and neurodegenerative disease but fall beyond the reach of the tools developed for classical structural biology due to their extreme structural flexibility.[2,3] Many IDPs undergo disorder-to-order transitions upon interaction with physiological partners, where molecular recognition is accompanied by local folding upon binding.[4,5] However the relationship between intrinsic conformational propensity and the structure adopted by the protein in its bound form remains poorly understood. For these reasons the development of meaningful descriptions of the conformational behavior of IDPs, and their relationship to protein function and malfunction, represents a key challenge for contemporary structural biology.

Nuclear magnetic resonance (NMR) spectroscopy reports on structural propensities at atomic resolution, on time scales varying over many orders of magnitude, and is therefore probably the most powerful biophysical tool for studying IDPs.[6] NMR inherently provides time- and ensemble-averaged structurally dependent experimental measurements and, as such, is exquisitely suited to the study of conformationally heterogeneous and flexible systems.[7] The dynamic averaging properties of NMR observables are well understood, rendering their exploitation particularly appropriate for the development of atomic resolution ensemble descriptions of flexible or unfolded proteins.[8−12]

Chemical shifts measured in IDPs report on the population weighted average over an entire ensemble of interchanging conformers, exchanging on time scales faster than the millisecond range. These readily measured parameters are nevertheless highly sensitive probes of the local protein conformation,[13−15] as has been demonstrated by the recent determination of three-dimensional structures of entire globular proteins using chemical shifts as sole experimental constraints.[16,17] The dependences of $^{13}C\alpha$ and $^{13}C\beta$ chemical shifts on backbone ϕ/ψ dihedral angles have been routinely used to identify the position of secondary structure and to estimate the level of secondary structural propensity within folded and unfolded proteins.[18−26] In this study we combine ensemble descriptions of unfolded proteins,[27] with a state-of-the-art chemical shift prediction algorithm that has underpinned the successful determination of folded proteins from chemical shifts.[28] This powerful combination is used to explore the possibility of using chemical shifts alone to map the local backbone conformational sampling of intrinsically disordered and partially folded proteins.

$^{13}C\alpha$, $^{13}C\beta$, $^{13}C'$, and ^{15}N chemical shifts exhibit different dependences on the backbone ϕ/ψ dihedral angles and are therefore sensitive probes of conformational sampling in disordered proteins.[29,30] These sensitivities are complementary in terms of the mapping of different regions of Ramachandran space, suggesting that their combination may allow the resolution of site-specific backbone conformational behavior. $^{13}C\alpha$ and $^{13}C\beta$ secondary shifts report essentially on the Ramachandran space sampled by the observed amino acid, while both $^{13}C'$ and ^{15}N are also sensitive to the sampling properties of the neighboring amino acids. To exploit this complementarity we employ an explicit ensemble description of unfolded proteins (*Flexible-Meccano*) that has been used in combination with residual dipolar coupling (RDCs),[28] scalar couplings,[31] and small angle scattering data[12,28] to describe conformational sampling in IDPs and chemically denatured proteins.

An efficient selection algorithm (ASTEROIDS)[32] is used to assemble a 200-strong subensemble of structures out of a much larger pool, which is in agreement with the experimental $^{13}C\alpha$, $^{13}C\beta$, $^{13}C'$, and ^{15}N chemical shifts. Selection starts from a large pool of conformers (typically 10 000 structures) constructed by *Flexible-Meccano* using standard random coil ϕ/ψ backbone dihedral angles. The program SPARTA is used to calculate chemical shifts for each member of the ensemble. The selection procedure involves two steps: an iteration step where each residue is treated independently, and a final step where full structures are selected. The first iteration step consists of the selection of 200 ϕ/ψ values for each residue that are in agreement with the $^{13}C\alpha$, $^{13}C\beta$, and $^{13}C'$ chemical shifts. This step is repeated five times to obtain 1000 ϕ/ψ values for each residue. A new ensemble of structures is created using *Flexible-Meccano*, but this time using the selected 1000 ϕ/ψ values for each residue. ASTEROIDS is applied again for each residue (see Supporting Information) independently to select 5×200 ϕ/ψ values from the new pool of structures. This iterative procedure is repeated until no further improvement in the fitting of the chemical shifts of the individual residues can be obtained. Step two of the selection procedure is then applied using $^{13}C\alpha$, $^{13}C\beta$, $^{13}C'$, and ^{15}N chemical shifts, where entire structures (200 conformers) are selected from the pool of structures generated during the previous iterations. While chemical shifts are expected to report only on local conformations, other experimental data such as residual dipolar couplings (RDCs) and paramagnetic relaxation enhancements (PREs) or SAXS report on long-range order. Therefore, the selection of entire structures will allow combined fitting of several types of experimental data.

No assumptions are made in terms of the secondary structural propensity, as the first ensemble contains only unfolded structures derived from the statistical coil database. Local conformational bias is recognized on the basis of chemical shift, and resulting propensities are used to assemble the new database for the subsequent iteration. In this way the algorithm automatically provides the appropriate backbone dihedral angles for the

 10.1021/ja909973n

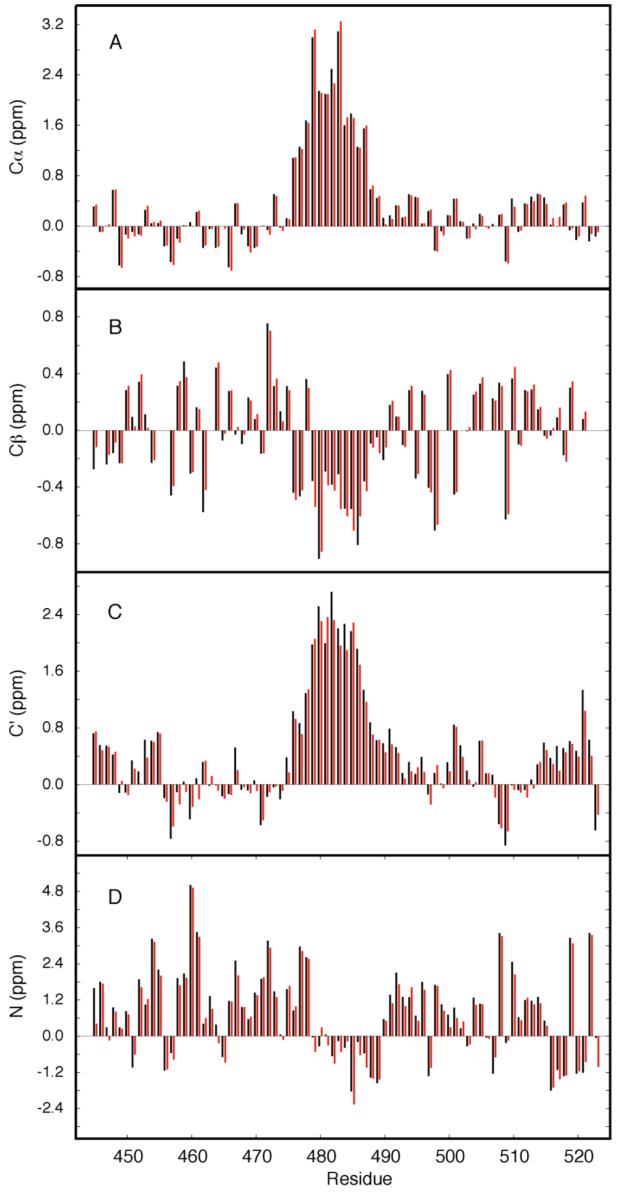

Figure 1. Reproduction of experimental secondary chemical shifts (random coil values from RefDB[25] were used) from an ensemble of 200 structures determined using the ASTEROIDS algorithm (3 iterations). Chemical shifts were calculated using the program SPARTA. *Flexible-Meccano* was used to calculate ensembles of structures of the protein, and iterative selection of 200-strong subensembles provided a final ensemble in agreement with experimental shifts. Black: experimental secondary chemical shifts. Red: secondary chemical shifts averaged over the final ensemble. (A) α carbon, (B) β carbon, (C) carbonyl, (D) amide nitrogen.

construction of entire secondary structural elements, as well as determining local conformational sampling in the unfolded domains.

This analysis is applied here to the study of N_{TAIL}, the C-terminal domain of the Sendai virus nucleoprotein. The molecular recognition element of N_{TAIL} has been shown, using detailed analysis of multiple RDCs, to contain a conformationally fluctuating helical element at its center.[33,34] This protein is a particularly appropriate model with which to test the approach, as it contains both partially structured and fully disordered elements. Figure 1 shows the agreement between the experimental and the calculated secondary chemical

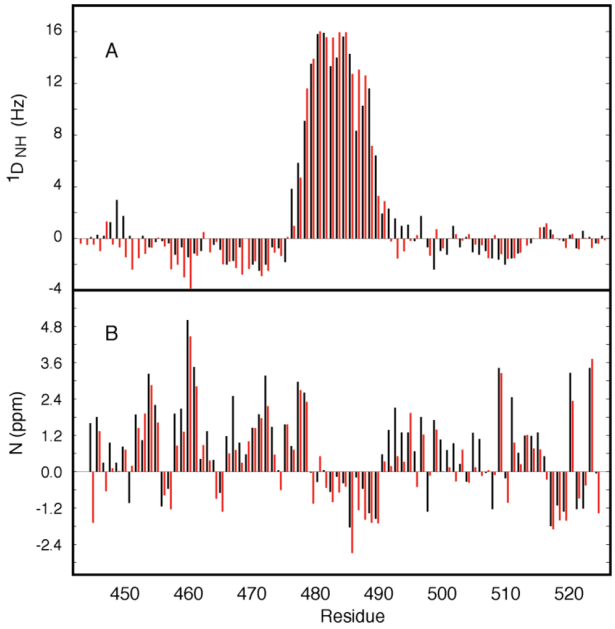

Figure 2. Reproduction of independent parameters by the ensemble based on chemical shift selection. (A) $^{15}N-^{1}H$ residual dipolar couplings (RDCs) measured in sterically aligned N_{TAIL}.[33] 50 000 conformers were calculated using the amino acid specific description of N_{TAIL} determined from the chemical shifts. RDCs were calculated using the program PALES directly from the ensemble and averaged. Simulated RDCs (red) were scaled uniformly to best match experiment (black). (B) Reproduction of ^{15}N secondary chemical shifts using an ensemble determined from only $^{13}C\alpha$, $^{13}C\beta$, $^{13}C'$ shifts (black: experiment, red: simulation).

shifts in N_{TAIL} after application of the ASTEROIDS algorithm. Excellent agreement with experimental shifts is observed throughout the protein.

The simple observation that data can be reproduced by a specific conformational ensemble does not necessarily guarantee that the ensemble is physically realistic. It is therefore essential to be able to cross-validate this approach with independent experimental data. Figure 2A shows the agreement between experimental $^{15}N-^{1}H$ RDCs measured from partially aligned N_{TAIL} compared to those calculated using an ensemble obtained from chemical shift derived conformational sampling. RDCs were calculated using the program PALES.[35] The agreement is striking, in both the folded and unfolded regions of the protein, demonstrating the ability of the algorithm to unambiguously interpret chemical shifts in terms of local conformational propensity. The level of helical structure agrees very closely with the helical description that was derived from analysis of RDCs,[28] indicating that the method is also quantitative. The ensemble dimensions also agree with those found in the previous studies (data not shown).[28] In a further test of consistency we have repeated the analysis in the absence of the ^{15}N chemical shifts and compared these shifts to predicted values (Figure 2B). Although this implies removing 25% of the data, experimental values are still reasonably reproduced (rmsd = 0.77 ppm compared to 1.15 ppm for the standard coil distribution).

An analysis of the ϕ/ψ distribution of the selected conformers outside the helical element reveals that the fully disordered regions of the protein have an overall tendency to sample less β-extended $\{\phi/\psi \approx -135°/135°\}$ and more (on average 5%) polyproline II $\{\phi/\psi \approx -75°/150°\}$ than is present in standard random coil

databases. These trends are in qualitative agreement with observations based on complementary spectroscopic techniques.[36−38] We are currently applying similar analyses to chemical shifts from more proteins to determine general trends for backbone conformational propensities of IDPs.

The ability of chemical shifts to reproduce conformational sampling was tested using extensive simulation. Ensembles of a model unfolded sequence were created using the standard ϕ/ψ database, an extended database sampling more β-sheet and polyproline II regions, or a database sampling more α-helical conformations. The chemical shifts of these ensembles were calculated with SPARTA as described above. The three sets of chemical shifts (standard, extended, and helix) were subjected to ASTEROIDS for selection of a subensemble of 200 structures from a pool of conformers generated using the standard database. These simulations demonstrate that it is possible to obtain a standard coil, more extended sampling, or a more helical sampling directly from chemical shifts, to within 5% accuracy (results not shown).

The ability to describe conformational sampling on the basis of chemical shifts alone is important for the development of atomic resolution descriptions of IDPs. The approach presented here makes no assumption concerning the true conformational properties of the molecule, starting with a standard statistical coil description of backbone conformational sampling, and refining this iteratively until convergence is reached compared to the experimental data. This allows the identification and characterization of entire secondary structural elements and their associated populations, as well as providing indications of the subtle detail of local conformational sampling in unfolded proteins. The approach is entirely compatible with recently presented ensemble selection algorithms based on the use of complementary structural restraints such as RDCs or ^3J scalar couplings, providing a tool for the development of a unified conformational model of partially ordered states. Possibly more exciting, this technique raises the prospect of probing the conformational behavior of unfolded proteins under conditions where additional parameters cannot be easily measured but where chemical shifts are still accessible, for example in crowded or cellular environments.[39]

Acknowledgment. This work was supported by the Commisariat à l'Energie Atomique, the French CNRS, the Université Joseph Fourier, Grenoble, and through ANR Protein Motion PCV (2007) ANR-07-PCVI PROTEIN MOTION. M.R.J. is supported by Lundbeckfonden and is recipient of a long-term EMBO fellowship.

Supporting Information Available: Complete ref 17. Materials and methods, describing use of SPARTA, application of ASTEROIDS for ensemble selection. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Uversky, V. N. *Protein Sci.* **2002**, *11*, 739–756.
(2) Dyson, H. J.; Wright, P. E. *Nat. Rev. Mol. Cell Biol.* **2005**, *6*, 197–208.
(3) Eliezer, D. *Curr. Opin. Struct. Biol.* **2009**, *19*, 23–30.
(4) Tompa, P.; Fuxreiter, M. *Trends Biochem. Sci.* **2008**, *33*, 2–8.
(5) Sugase, K.; Dyson, H. J.; Wright, P. E. *Nature* **2007**, *447*, 1021–1025.
(6) Dyson, H. J.; Wright, P. E. *Chem. Rev.* **2004**, *104*, 3607–3622.
(7) Mittag, T.; Forman-Kay, J. D. *Curr. Opin. Struct. Biol.* **2007**, *17*, 3–14.
(8) Jha, A. K.; Colubri, A.; Freed, K.; Sosnick, T. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 13099–13105.
(9) Marsh, J. A.; Forman-Kay, J. D. *J. Mol. Biol.* **2009**, *391*, 359–374.
(10) Jensen, M. R.; Markwick, P.; Griesinger, C.; Zweckstetter, M.; Meier, S.; Grzesiek, S.; Bernado, P.; Blackledge, M. *Structure* **2009**, *17*, 1169–1185.
(11) Wells, M.; Tidow, H.; Rutherford, T. J.; Markwick, P.; Jensen, M. R.; Mylonas, E.; Svergun, D. I.; Blackledge, M.; Fersht, A. R. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 5762–5767.
(12) Krzeminski, M.; Fuentes, G.; Boelens, R.; Bonvin, A. M. J. J. *Proteins* **2009**, *74*, 895–904.
(13) Wishart, D. S.; Sykes, B. D.; Richards, F. M. *J. Mol. Biol.* **1991**, *222*, 311–333.
(14) Spera, S.; Bax, A. *J. Am. Chem. Soc.* **1991**, *113*, 5490–5492.
(15) Osapay, K.; Case, D. A. *J. Biomol. NMR* **1994**, *4*, 215–230.
(16) Cavalli, A.; Salvatella, X.; Dobson, C. M.; Vendruscolo, M. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 9615–9620.
(17) Shen, Y.; et al. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 4685–4690.
(18) Schwarzinger, S.; Kroon, G. J. A.; Foss, T. R.; Chung, J.; Wright, P. E.; Dyson, H. J. *J. Am. Chem. Soc.* **2001**, *123*, 2970–2978.
(19) Yao, J.; Chung, J.; Eliezer, D.; Wright, P. E.; Dyson, H. J. *Biochemistry.* **2001**, *40*, 3561–3571.
(20) Wang, Y. J.; Jardetzky, O. *J. Am. Chem. Soc.* **2002**, *124*, 14075–14084.
(21) Eghbalnia, H. R.; Wang, L.; Bahrami, A.; Assadi, A.; Markley, J. L. *J. Biomol. NMR* **2005**, *32*, 71–81.
(22) Marsh, J. A.; Singh, V. K.; Jia, Z. C.; Forman-Kay, J. D. *Protein Sci.* **2006**, *15*, 2795–2804.
(23) Modig, K.; Jürgensen, V. W.; Lindorff-Larsen, K.; Fieber, W.; Bohr, H. G.; Poulsen, F. M. *FEBS Lett.* **2007**, *581*, 4965–4971.
(24) De Simone, A.; Cavalli, A.; Hsu, S. T.; Vranken, W.; Vendruscolo, M. *J. Am. Chem. Soc.* **2009**, *131*, 16332–16333.
(25) Zhang, H. Y.; Neal, S.; Wishart, D. S. *J. Biomol. NMR* **2003**, *25*, 173–195.
(26) Peti, W.; Smith, L. J.; Redfield, C.; Schwalbe, H. *J. Biomol. NMR* **2001**, *19*, 153–165.
(27) Bernado, P.; Blanchard, L.; Timmins, P.; Marion, D.; Ruigrok, R. W. H.; Blackledge, M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 17002–17007.
(28) Shen, Y.; Bax, A. *J. Biomol. NMR* **2007**, *38*, 289–302.
(29) Alexandrescu, A. T.; Abeygunawardana, C.; Shortle, D. *Biochemistry* **1994**, *33*, 1063–1072.
(30) Zhang, O. W.; Kay, L. E.; Olivier, J. P.; Forman-Kay, J. D. *J. Biomol. NMR* **1994**, *4*, 845–858.
(31) Meier, S.; Grzesiek, S.; Blackledge, M. *J. Am. Chem. Soc.* **2007**, *129*, 9799–9807.
(32) Nodet, G.; Salmon, L.; Ozenne, V.; Meier, S.; Jensen, M. R.; Blackledge, M. *J. Am. Chem. Soc.* **2009**, *131*, 16968–16975.
(33) Jensen, M. R.; Houben, K.; Lescop, E.; Blanchard, L.; Ruigrok, R. W. H.; Blackledge, M. *J. Am. Chem. Soc.* **2008**, *130*, 8055–8061.
(34) Jensen, M. R.; Blackledge, M. *J. Am. Chem. Soc.* **2008**, *130*, 11266–11267.
(35) Zweckstetter, M.; Bax, A. *J. Am. Chem. Soc.* **2000**, *122*, 3791–3792.
(36) Maiti, N. C.; Apetri, M. M.; Zagorski, M. G.; Carey, P. R.; Anderson, V. E. *J. Am. Chem. Soc.* **2004**, *126*, 2399–2408.
(37) Shi, Z. S.; Chen, K.; Liu, Z. G.; Kallenbach, N. R. *Chem. Rev.* **2006**, *106*, 1877–1897.
(38) Woody, R. W. *J. Am. Chem. Soc.* **2009**, *131*, 8234–8245.
(39) Serber, Z.; Dötsch, V. *Biochemistry* **2001**, *40*, 14317–14323.

JA909973N

J|A|C|S
ARTICLES

Published on Web 05/25/2010

NMR Characterization of Long-Range Order in Intrinsically Disordered Proteins

Loïc Salmon,[†,§] Gabrielle Nodet,[†,§] Valéry Ozenne,[†] Guowei Yin,[‡] Malene Ringkjøbing Jensen,[†] Markus Zweckstetter,[‡] and Martin Blackledge*[,†]

Protein Dynamics and Flexibility, Institut de Biologie Structurale Jean-Pierre Ebel, CEA; CNRS; UJF UMR 5075, 41 Rue Jules Horowitz, Grenoble 38027, France, and NMR-Based Structural Biology, Max Planck Institute for Biophysical Chemistry, 37077 Göttingen, Germany

Received February 25, 2010; E-mail: martin.blackledge@ibs.fr

Abstract: Intrinsically disordered proteins (IDPs) are predicted to represent a significant fraction of the human genome, and the development of meaningful molecular descriptions of these proteins remains a key challenge for contemporary structural biology. In order to describe the conformational behavior of IDPs, a molecular representation of the disordered state based on diverse sources of structural data that often exhibit complex and very different averaging behavior is required. In this study, we propose a combination of paramagnetic relaxation enhancements (PREs) and residual dipolar couplings (RDCs) to define both long-range and local structural features of IDPs in solution. We demonstrate that ASTEROIDS, an ensemble selection algorithm, faithfully reproduces intramolecular contacts, even in the presence of highly diffuse, ill-defined target interactions. We also show that explicit modeling of spin-label mobility significantly improves the reproduction of experimental PRE data, even in the case of highly disordered proteins. Prediction of the effects of transient long-range contacts on RDC profiles reveals that weak intramolecular interactions can induce a severe distortion of the profiles that compromises the description of local conformational sampling if it is not correctly taken into account. We have developed a solution to this problem that involves efficiently combining RDC and PRE data to simultaneously determine long-range and local structure in highly flexible proteins. This combined analysis is shown to be essential for the accurate interpretation of experimental data from α-synuclein, an important IDP involved in human neurodegenerative disease, confirming the presence of long-range order between distant regions in the protein.

Introduction

The realization that a large fraction of functional proteins encoded by the human genome are intrinsically disordered or contain long disordered regions has revealed a fundamental limitation of classical structural biology.[1−4] Intrinsically disordered proteins (IDPs) are functional despite their lack of well-defined structure, imposing a new perspective on the relationship between primary protein sequence and function and necessitating the development of an entirely new set of experimental and analytical techniques.[5,6] The importance of developing new methodologies to study these proteins is underlined by the fact that IDPs are associated with many human diseases, including cancer, cardiovascular disease, amyloidosis, neurodegenerative disease, and diabetes.

NMR spectroscopy is exquisitely suited to the study of IDPs,[7] primarily because heteronuclear chemical shift assignment remains possible even for very large disordered proteins.[8] NMR analysis can then be used to precisely study the specific local conformational preferences that encode biological function.[9−11] In spite of their highly dynamic nature, IDPs also exhibit transient or persistent long-range tertiary structure that may be related to biological activity (e.g., via so-called fly-casting interactions[12]) or simply confer protection from proteolysis or amyloidosis. It is precisely the transient nature of such contacts that precludes straightforward NMR detection using standard techniques such as $^{1}H-^{1}H$ cross-relaxation. However, long-range information can be measured via the effects of dipolar relaxation between the observed spin and an unpaired electron, which can be artificially introduced into the protein by attaching a nitroxide group to a strategically placed cysteine mutant.[13,14]

[†] Institut de Biologie Structurale Jean-Pierre Ebel.
[‡] Max Planck Institute for Biophysical Chemistry.
[§] These authors contributed equally.

(1) Uversky, V. N. *Protein Sci.* **2002**, *11*, 739–756.
(2) Dunker, A. K.; Brown, C. J.; Lawson, J. D.; Iakoucheva, L. M.; Obradovic, Z. *Biochemistry* **2002**, *41*, 6573–6582.
(3) Tompa, P. *Trends. Biochem. Sci.* **2002**, *27*, 527–533.
(4) Dyson, H. J.; Wright, P. E. *Curr. Opin. Struct. Biol.* **2002**, *12*, 54–60.
(5) Mittag, T.; Forman-Kay, J. D. *Curr. Opin. Struct. Biol.* **2007**, *17*, 3–14.
(6) Eliezer, D. *Curr. Opin. Struct. Biol.* **2009**, *19*, 23–30.
(7) Dyson, H. J.; Wright, P. E. *Chem. Rev.* **2004**, *104*, 3607–3622.
(8) Mukrasch, M. D.; Bibow, S.; Korukottu, J.; Jeganathan, S.; Biernat, J.; Griesinger, C.; Mandelkow, E. M.; Zweckstetter, M. *PLoS Biol.* **2009**, *7*, 399–414.
(9) Meier, S.; Blackledge, M.; Grzesiek, S. *J. Chem. Phys.* **2008**, *128*, 052204.
(10) Wright, P. E.; Dyson, H. J. *Curr. Opin. Struct. Biol.* **2009**, *19*, 31–38.
(11) Jensen, M. R.; Markwick, P.; Griesinger, C.; Zweckstetter, M.; Meier, S.; Grzesiek, S.; Bernado, P.; Blackledge, M. *Structure* **2009**, *17*, 1169–1185.
(12) Shoemaker, B. A.; Portman, J. J.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 8868–8873.
(13) Gillespie, J. R.; Shortle, D. *J. Mol. Biol.* **1997**, *268*, 158–169.
(14) Clore, G. M.; Tang, C.; Iwahara, J. *Curr. Opin. Struct. Biol.* **2007**, *17*, 603–616.

The gyromagnetic ratio of the electron spin is sufficiently high that the observed line broadening due to the paramagnetic relaxation enhancement (PRE) affords sensitive long-range probes of intra- and intermolecular distances and distance distribution functions. The interpretation of experimental PREs can be relatively straightforward in the case of folded proteins, where an assumption of a static probe localized at a single point in space can be applied to extract approximate distance constraints.[15] It has also been shown that simple modeling of spin-label side-chain mobility in terms of an average over three positions can significantly improve the accuracy of the distance information.[16] Detailed information about transient encounter complexes and their role in protein−protein interactions can also be extracted by combining paramagnetic effects and ensemble-averaged restrained molecular dynamics (MD).[17−19]

In the case of partially folded and unfolded proteins, paramagnetic effects are particularly powerful, as the interactions are sufficiently strong to allow the identification of fluctuating, weakly populated tertiary structural contacts. In this case, the treatment of the intrinsic dynamics of the system is of considerable importance. PREs have thus been interpreted in terms of average distance restraints between the unpaired electron and the observed spin, and these distances have been incorporated directly as constraints into restrained MD or ensemble-averaged restrained MD approaches.[20−27] Explicit relaxation rates can also be incorporated as constraints,[28] and more recently, PREs have been interpreted in terms of probability distributions.[26,29,30] PREs can also be used to select representative ensembles from a large pool of possible conformers.[31−33]

In this study, we have applied to the interpretation of PRE data from disordered proteins a recently introduced approach for modeling highly dynamic and disordered systems that derives explicit molecular ensembles on the basis of experimental data. Ensemble selection is based on the creation of a large number of conformers using an amino acid-specific random coil database known as *flexible-meccano*.[34] *Flexible-meccano* allows for very efficient restraint-free sampling of the available conformational space and was initially demonstrated and refined to provide structural ensembles in agreement with experimentally measured NMR and small-angle X-ray scattering (SAXS) data.[35−42] In parallel, the ensemble selection algorithm ASTEROIDS has been developed to directly determine appropriate regions of conformational space populated by the IDP through selection of conformers from the *flexible-meccano* ensemble using inferential analysis of experimental NMR data.[43] To date, the approach has been applied to experimental measurements that depend essentially on local structural behavior, such as residual dipolar couplings (RDCs) and chemical shifts.[44] Here we have adapted the approach to incorporate the interpretation of PREs. In order to allow for flexibility of the spin label with respect to the backbone conformation, explicit rotameric libraries that have been parametrized against experimental electron spin resonance (ESR) measurements and MD simulations[45] are used to map the allowed position of the electron spin. We then account for the dynamics of the electron spin within this envelope by evoking a model for the autocorrelation function of the relaxation-active interaction that was originally proposed for the interpretation of $^1H-^1H$ cross-relaxation effects.[46] This allows the motion of the relaxation-active dipole−dipole interaction between the electron spin and the observed nucleus to be modeled for each conformer in the ensemble.

The observation that RDCs can be measured in disordered proteins has been followed by the rapid development of techniques for interpreting experimental data in terms of local structure.[38,40,41,47−60] Comparison of experimental data with

(15) Battiste, J. L.; Wagner, G. *Biochemistry* **2000**, *39*, 5355–5365.
(16) Iwahara, J.; Schwieters, C. D.; Clore, G. M. *J. Am. Chem. Soc.* **2004**, *126*, 5879–5896.
(17) Tang, C.; Schwieters, C. D.; Clore, G. M. *Nature* **2007**, *449*, 1078–1082.
(18) Volkov, A. N.; Worrall, J. A.; Holtzmann, E.; Ubbink, M. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 18945–18950.
(19) Bashir, Q.; Volkov, A. N.; Ullmann, G. M.; Ubbink, M. *J. Am. Chem. Soc.* **2010**, *132*, 241–247.
(20) Gillespie, J. R.; Shortle, D. *J. Mol. Biol.* **1997**, *268*, 170–184.
(21) Lindorff-Larsen, K.; Kristjansdottir, S.; Teilum, K.; Fieber, W.; Dobson, C. M.; Poulsen, F. M.; Vendruscolo, M. *J. Am. Chem. Soc.* **2004**, *126*, 3291–3299.
(22) Dedmon, M. M.; Lindorff-Larsen, K.; Christodoulou, J.; Vendruscolo, M.; Dobson, C. M. *J. Am. Chem. Soc.* **2005**, *127*, 476–477.
(23) Bertoncini, C. W.; Jung, Y. S.; Fernandez, C. O.; Hoyer, W.; Griesinger, C.; Jovin, T. M.; Zweckstetter, M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 1430–1435.
(24) Kristjansdottir, S.; Lindorff-Larsen, K.; Fieber, W.; Dobson, C. M.; Vendruscolo, M.; Poulsen, F. M. *J. Mol. Biol.* **2005**, *347*, 1053–1062.
(25) Song, J.; Guo, L. W.; Muradov, H.; Artemyev, N. O.; Ruoho, A. E.; Markley, J. L. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 1505–1510.
(26) Allison, J. R.; Varnai, P.; Dobson, C. M.; Vendruscolo, M. *J. Am. Chem. Soc.* **2009**, *131*, 18314–18326.
(27) Ganguly, D.; Chen, J. *J. Mol. Biol.* **2009**, *390*, 467–477.
(28) Huang, J.-R.; Grzesiek, S. *J. Am. Chem. Soc.* **2010**, *132*, 694–705.
(29) Felitsky, D. J.; Lietzow, M. A.; Dyson, H. J.; Wright, P. E. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 6278–6283.
(30) Xue, Y.; Podkorytov, I. S.; Rao, D. K.; Benjamin, N.; Sun, H.; Skrynnikov, N. R. *Protein Sci.* **2009**, *18*, 1401–1424.
(31) Marsh, J. A.; Neale, C.; Jack, F. E.; Choy, W.-Y.; Lee, A. Y.; Crowhurst, K. A.; Forman-Kay, J. D. *J. Mol. Biol.* **2007**, *367*, 1494–1510.
(32) Marsh, J. A.; Forman-Kay, J. D. *J. Mol. Biol.* **2009**, *391*, 359–374.
(33) Cho, M. K.; Nodet, G.; Kim, H. Y.; Jensen, M. R.; Bernado, P.; Fernandez, C. O.; Becker, S.; Blackledge, M.; Zweckstetter, M. *Protein Sci.* **2009**, *18*, 1840–1846.
(34) Bernado, P.; Blanchard, L.; Timmins, P.; Marion, D.; Ruigrok, R. W. H.; Blackledge, M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 17002–17007.
(35) Bernado, P.; Bertoncini, C.; Griesinger, C.; Zweckstetter, M.; Blackledge, M. *J. Am. Chem. Soc.* **2005**, *127*, 17968–17969.
(36) Dames, S. A.; Aregger, R.; Vajpai, N.; Bernado, P.; Blackledge, M.; Grzesiek, S. *J. Am. Chem. Soc.* **2006**, *128*, 13508–13514.
(37) Skora, L.; Cho, M. K.; Kim, H.-Y.; Fernandez, C.; Blackledge, M.; Zweckstetter, M. *Angew. Chem., Int. Ed.* **2006**, *45*, 7012–7015.
(38) Mukrasch, M. D.; Markwick, P. R. L.; Biernat, J.; von Bergen, M.; Bernado, P.; Griesinger, C.; Mandelkow, E.; Zweckstetter, M.; Blackledge, M. *J. Am. Chem. Soc.* **2007**, *129*, 5235–5243.
(39) Meier, S.; Grzesiek, S.; Blackledge, M. *J. Am. Chem. Soc.* **2007**, *129*, 9799–9807.
(40) Jensen, M. R.; Houben, K.; Lescop, E.; Blanchard, L.; Ruigrok, R. W. H.; Blackledge, M. *J. Am. Chem. Soc.* **2008**, *130*, 8055–8061.
(41) Wells, M.; Tidow, H.; Rutherford, T. J.; Markwick, P.; Jensen, M. R.; Mylonas, E.; Svergun, D. I.; Blackledge, M.; Fersht, A. R. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 5762–5767.
(42) Bernado, P.; Blackledge, M. *Biophys. J.* **2009**, *97*, 2839–2845.
(43) Nodet, G.; Salmon, L.; Ozenne, V.; Meier, S.; Jensen, M. R.; Blackledge, M. *J. Am. Chem. Soc.* **2009**, *131*, 16968–16975.
(44) Jensen, M. R.; Salmon, L.; Nodet, G.; Blackledge, M. *J. Am. Chem. Soc.* **2010**, *132*, 1270–1272.
(45) Sezer, D.; Freed, J. H.; Roux, B. *J. Phys. Chem. B* **2008**, *112*, 5755–5767.
(46) Brüschweiler, R.; Roux, B.; Blackledge, M.; Griesinger, C.; Karplus, M.; Ernst, R. R. *J. Am. Chem. Soc.* **1992**, *114*, 2289–2302.
(47) Shortle, D.; Ackerman, M. S. *Science* **2001**, *293*, 487–489.
(48) Alexandrescu, A. T.; Kammerer, R. A. *Protein Sci.* **2003**, *12*, 2132–2140.
(49) Mohana-Borges, R.; Goto, N. K.; Kroon, G. J. A.; Dyson, H. J.; Wright, P. E. *J. Mol. Biol.* **2004**, *340*, 1131–1142.
(50) Fieber, W.; Kristjansdottir, S.; Poulsen, F. M. *J. Mol. Biol.* **2004**, *339*, 1191–1199.
(51) Meier, S.; Güthe, S.; Kiefhaber, T.; Grzesiek, S. *J. Mol. Biol.* **2004**, *344*, 1051–1069.

predictions from calculated ensembles of random-coil conformers has indicated that RDCs are sensitive to amino acid-specific backbone dihedral angle distributions. The ability to define random-coil RDC values has led to first the identification and then the quantification of the level of secondary structure propensity in IDPs, initially by comparison with ensemble averages reporting on different sampling regimes[35−42] and more recently by using RDCs to determine conformational sampling on an amino acid-specific basis using ASTEROIDS.[43] In the latter case, a highly efficient local alignment window (LAW) approach to the simulation of RDCs was used to account for local-sampling and near-neighbor effects.[43,59] This demonstrated that in order to correctly define the conformational behavior for a LAW with a length of 15 amino acids, at least 200 structures are needed to average the RDCs.[43] In addition, it was noted that in contrast to chemical shifts and scalar couplings, RDCs are also sensitive to the degree and nature of transient long-range order, and even in the absence of specific contacts, it was found to be necessary to combine the local prediction from the LAWs with a generic baseline profile along the primary sequence that accounts for the chainlike nature of the protein.

In this study, ASTEROIDS and *flexible-meccano* were adapted to allow for transient long-range order and combined with experimental PREs to determine an ensemble description of α-synuclein, a paradigm of the IDP family, whose conformational properties in free solution have been characterized extensively using NMR spectroscopy and associated biophysical techniques.[22,23,26,61−66] We demonstrate that even in the presence of highly diffuse, ill-defined target interactions, explicit modeling of spin-label mobility significantly improves the prediction of experimental data not used in the analysis. We also show that even weak intramolecular interactions can induce a severe distortion of the expected RDC values that compromises the description of local conformational sampling if not correctly taken into account. The expected modulation of the RDCs is parametrized in such a way that it can be analytically introduced into the predicted RDC profile, and we demonstrate that incorporation of long-range contacts from the PRE-derived ensemble significantly improves the prediction of experimental RDCs from α-synuclein.[23] This novel approach allows for the direct and efficient introduction of long-range contacts into ensemble-averaged RDCs and provides for the simple and powerful combination of RDCs and PREs into a single ensemble description.

Theoretical Aspects

Dynamic Averaging of PREs. IDPs are highly flexible on diverse time scales, and this flexibility must be taken into account in the analysis of the measured PREs. The transverse relaxation rate due to the presence of the unpaired electron, Γ_2, can be expressed as follows:[67]

$$\Gamma_2 = \frac{2}{5}\left(\frac{\mu_0}{4\pi}\right)^2 \gamma_H{}^2 g_e{}^2 \mu_B{}^2 s_e(s_e + 1)[4J(0) + 3J(\omega_H)] \quad (1)$$

where g_e is the electron g-factor, γ_H is the gyromagnetic ratio of the observed nucleus (proton), s_e is the electron spin, ω_H is the proton frequency, μ_B is the Bohr magneton, and μ_0 is the permittivity of free space. It has been shown[14,46] that the spectral density function $J(\omega)$ can be described using a model-free expression of the order parameter comprising the orientational and distance-dependent components of the internal motion, both of which strongly depend on the motion of the spin label with respect to the observed nuclear spin:

$$J(\omega) = \langle r_{H-e}{}^{-6} \rangle \left[\frac{S^2_{H-e}\tau_c}{1 + \omega^2\tau_c{}^2} + \frac{(1 - S^2_{H-e})\tau_e}{1 + \omega^2\tau_e{}^2} \right] \quad (2)$$

where the order parameter S^2_{H-e} describes the motion of the dipolar interaction vector, $\tau_c = \tau_r\tau_s/(\tau_r + \tau_s)$ is defined in terms of the electron spin and rotational correlation times τ_s and τ_r, respectively, τ_e is given by the expression $\tau_e = 1/(\tau_i{}^{-1} + \tau_r{}^{-1} + \tau_s{}^{-1})$, in which τ_i represents the effective correlation time of the spin label, and r_{H-e} is the instantaneous distance between the proton and electron spins. The order parameter can be usefully decomposed into radial and angular components as

$$S^2_{H-e} = S^2_{ang} S^2_{rad} \quad (3a)$$

where

$$S^2_{rad} = \langle r_{H-e}{}^{-6} \rangle^{-1} \langle r_{H-e}{}^{-3} \rangle^2 \quad (3b)$$

and

$$S^2_{ang} = \frac{4\pi}{5} \sum_{m=-2}^{2} |\langle Y_2^m(\Omega^{mol}) \rangle|^2 \quad (3c)$$

in which Ω^{mol} refers to the orientation of the interaction vector in the frame of the *flexible-meccano* conformer. These expressions are used to calculate the effective transverse relaxation rate for each backbone conformation produced with the *flexible-meccano* algorithm.

The electron spin label is attached to the molecule via a thiol-reactive methanethiosulfonate (MTSL) attached to a cysteine side chain. MTSL conformations are built explicitly for each *flexible-meccano* backbone conformer by randomly sampling known rotameric descriptions.[45] Only conformations that do not result in steric overlap with the remainder of the chain are retained in the *N*-conformer ensemble that is used to represent the position of the side chain. Thus, for each backbone

(52) Ohnishi, S.; Lee, A. L.; Edgell, M. H.; Shortle, D. *Biochemistry* **2004**, *43*, 4064–4070.
(53) Sallum, C. O.; Martel, D. M.; Fournier, R. S.; Matousek, W. M.; Alexandrescu, A. T. *Biochemistry* **2005**, *44*, 6392–6403.
(54) Ding, K.; Louis, J. M.; Gronenborn, A. M. *J. Mol. Biol.* **2004**, *335*, 1299–1307.
(55) Louhivuori, M.; Pääkkönen, K.; Fredriksson, K.; Permi, P.; Lounila, J.; Annila, A. *J. Am. Chem. Soc.* **2003**, *125*, 15647–15650.
(56) Jha, A. K.; Colubri, A.; Freed, K.; Sosnick, T. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 13099–13105.
(57) Obolensky, O. I.; Schlepckow, K.; Schwalbe, H.; Solov'yov, A. V. *J. Biomol. NMR* **2007**, *39*, 1–16.
(58) Betancourt, M. R. *J. Phys. Chem. B* **2008**, *112*, 5058–5069.
(59) Marsh, J. A.; Baker, J. M. R.; Tollinger, M.; Forman-Kay, J. D. *J. Am. Chem. Soc.* **2008**, *130*, 7804–7805.
(60) Jensen, M. R.; Blackledge, M. *J. Am. Chem. Soc.* **2008**, *130*, 11266–11267.
(61) Eliezer, D.; Kutluay, E.; Bussell, R., Jr.; Browne, G. *J. Mol. Biol.* **2001**, *307*, 1061–1073.
(62) Fernandez, C. O.; Hoyer, W.; Zweckstetter, M.; Jares-Erijman, E. A.; Subramaniam, V.; Griesinger, C.; Jovin, T. M. *EMBO J.* **2004**, *23*, 2039–2046.
(63) Sung, Y.-h.; Eliezer, D. *J. Mol. Biol.* **2007**, *372*, 689–707.
(64) Wu, K. P.; Kim, S.; Fela, D. A.; Baum, J. *J. Mol. Biol.* **2008**, *378*, 1104–1115.
(65) Li, C.; Lutz, E. A.; Slade, K. M.; Ruf, R. A.; Wang, G. F.; Pielak, G. J. *Biochemistry* **2009**, *48*, 8578–8584.
(66) Lendel, C.; Damberg, P. *J. Biomol. NMR* **2009**, *44*, 35–42.
(67) Solomon, I. *Phys. Rev.* **1955**, *99*, 559–565.

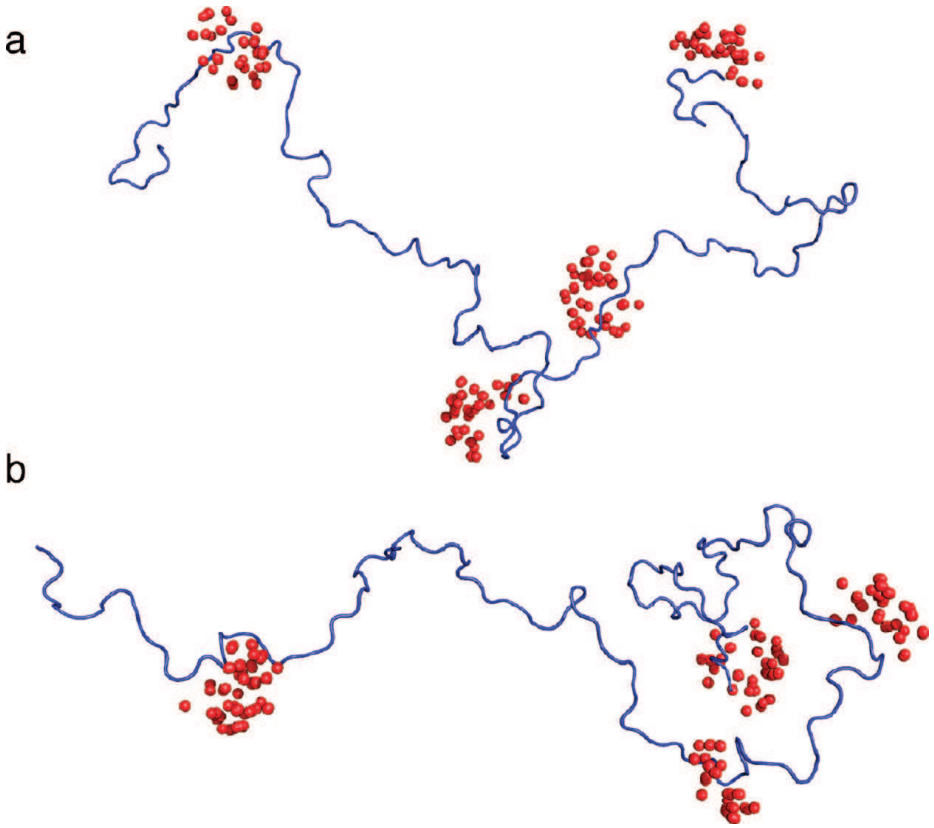

Figure 1. Representation of the possible nitroxide spin label positions relative to the backbone of individual structures calculated using the conformational sampling algorithm *flexible-meccano*. Two representative conformers are shown. The positions of the heavy atoms are represented by the blue ribbon, while allowed MTSL side-chain positions are shown in red for each of four paramagnetic probes used in the α-synuclein study (amino acids 18, 76, 90, and 140). Previously proposed MTSL rotameric libraries[45] were randomly sampled for a total of 600 conformers for each site. Each position was retained and included in the averaging procedure if no steric clashes were found with respect to the given backbone conformation.

conformation, the MTSL side chain is represented by a population-weighted sampling of the available rotameric states. The effective relaxation rate for each amide proton is taken as the average of the rates $\Gamma_{2,c}^{fm}$ for the N retained *flexible-meccano* conformers:

$$\Gamma_2^{\text{total}} = \frac{1}{N}\sum_{c=1}^{N}\Gamma_{2,c}^{fm} \tag{4}$$

Effective intensities are then calculated as described in Methods.

The assumption made here are that the interconversion between different side-chain conformations is independent of (and faster than) the interconversion between different discrete conformers. In common with previous applications,[23,28] we estimated τ_c to be 5 ns, and the internal motion describing the sampling of the different side-chain conformations was assumed to have a correlation time of 500 ps. This is in broad agreement with values derived from earlier MD/ESR-based studies,[68] and we note that changing the internal correlation time by a factor of 2 in either direction had no noticeable influence on the resulting analysis.

Figure 1 shows the possible positions of the spin label for each of four paramagnetic probes attached to cysteine mutants of the protein α-synuclein in two *flexible-meccano* conformers (amino acids 18, 76, 90, and 140, which are the positions used in the experimental study).[23] The spin label can clearly occupy a large volume space that could potentially affect the effective relaxation behavior of the observed spins.

Results and Discussion

Our aim in this study was to analyze the effects of long-range transient contacts on experimentally observable NMR parameters from unfolded proteins and to develop a formalism that allows their use for the meaningful characterization of both local and long-range structure in these highly flexible systems. In order to do this, we initially used molecular simulations to investigate the expected effects in systems with either one or two dominant long-range contacts. Although these simulated systems were intentionally oversimplified for the sake of clarity, the application of the observed results to more complex networks of long-range transient interactions is expected to be straightforward.

Paramagnetic Relaxation Enhancement in Highly Disordered Systems: Simulation. We initially determined whether it is possible to detect weakly specific long-range interactions via the combined ASTEROIDS and *flexible-meccano* analysis applied to simulated PREs. Figure 2 shows PREs calculated for a simulated model protein of 100 amino acids with paramagnetic spin labels attached at positions 20, 40, 60, and 80 (red bars). In Figure 2a, each conformer contains a contact between 41−50 and 81−90. The definition of a contact is given in Methods. The solid line shows the expected broadening in the absence of *specific* contacts (the reference ensemble where all conformers are allowed). We note that the effective broadening, even in the absence of specific contacts, is quite significant

(68) Sezer, D.; Freed, J. H.; Roux, B. *J. Chem. Phys.* **2008**, *128*, 165106.

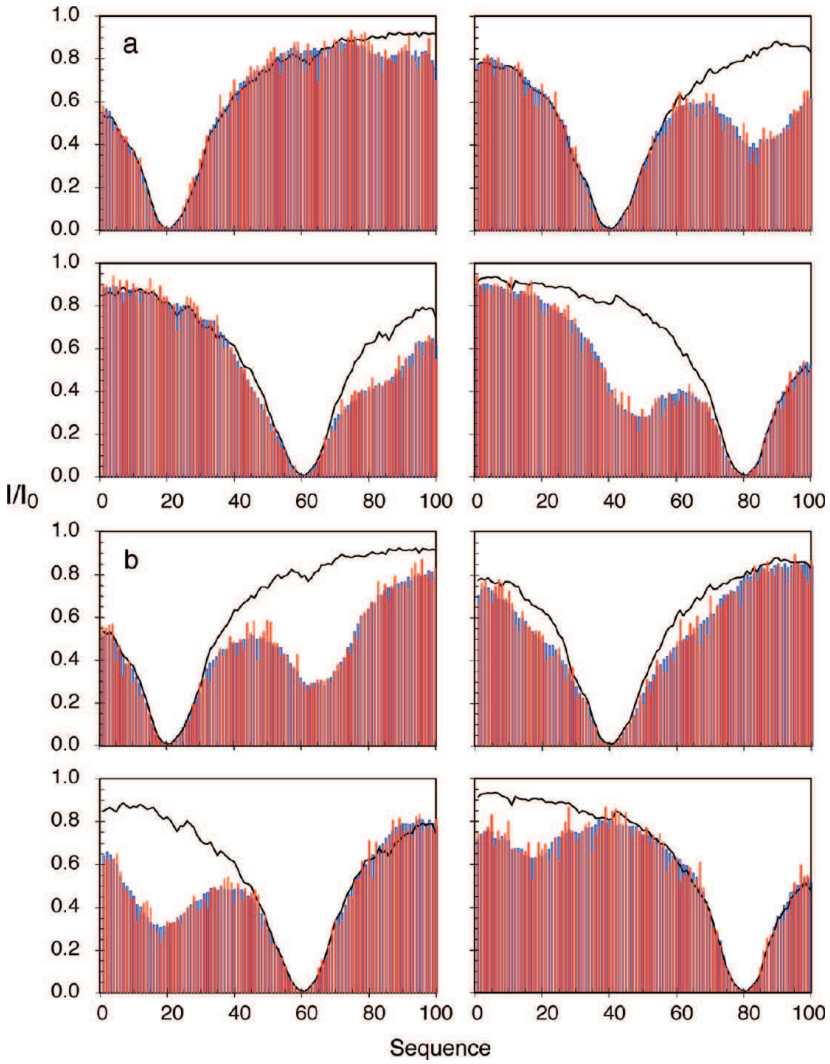

Figure 2. Reproduction of simulated sample PRE data for ensembles containing specific contacts using the ASTEROIDS ensemble selection algorithm.[43] (a) Blue: data averaged over the target ensemble in which each conformer has a contact between 41−50 and 81−90. Red: data averaged over an ensemble of 80 structures selected using ASTEROIDS. The four boxes show the PRE data for simulated spin labels at residues 20 (top left), 40 (top right), 60 (bottom left), and 80 (bottom right). Lines show the PREs calculated from a control ensemble with no specific contacts. (b) Blue: as in (a) for a target ensemble in which each conformer has a contact between 11−20 and 61−70. Red: data averaged over an ensemble of 80 structures selected using ASTEROIDS.

as a result of the large volume space sampled by the spin label. Figure 2b shows a similar representation of an ensemble with contacts between positions 11−20 and 61−70. The ASTEROIDS algorithm targeting these simulated PREs was then used to select 80-member conformational ensembles from a pool of 10 000 structures without specific contacts calculated using the *flexible-meccano* Monte Carlo sampling approach (see Methods). The resulting ensembles reproduced the simulated PREs well, as shown by the blue bars in Figure 2. It should be noted that these simulations used examples that were quite demanding, with 20% of the chain involved in weakly specific contacts. These simulations nevertheless represent a reasonable reproduction of the situation that one may encounter when studying intrinsically disordered or partially folded proteins, with long-range interactions occurring between strands carrying complementary electrostatic charge or containing hydrophobic side-chains. It was therefore of interest to determine whether the broad averaging effects predicted from such a simulation would allow the extraction of meaningful information concerning the long-range contacts.

ASTEROIDS Reproduces the Overall Biophysical Features of the Target Ensemble. Figure 3 shows the effective contacts present in the ASTEROIDS ensembles that matched the simulated data. This representation compares interatomic (C^{α}) distances present in the reference ensemble with those in the selected ensemble (see Methods). The contacts that were used to simulate the data are well identified in both cases. The exact values of the distances were not reproduced (the distances were underestimated), but this is not considered a serious drawback in view of the ill-defined nature of the contact. We also compared the overall distributions of the selected ensembles relative to the reference ensemble. Figure 4 shows that the ASTEROIDS ensemble of structures selected using the simulated PREs from the ensemble containing contacts between regions 11−20 and 61−70 (Figure 2b) reproduced the distribution of the radiii of gyration (R_g) for members of the target ensemble quite closely. The average R_g of the ASTEROIDS ensembles increased slightly with increasing number of structures, from 21.3 Å for the 80-member ensemble to 21.7 Å for the 160-member ensemble, compared with 22.6 Å for the

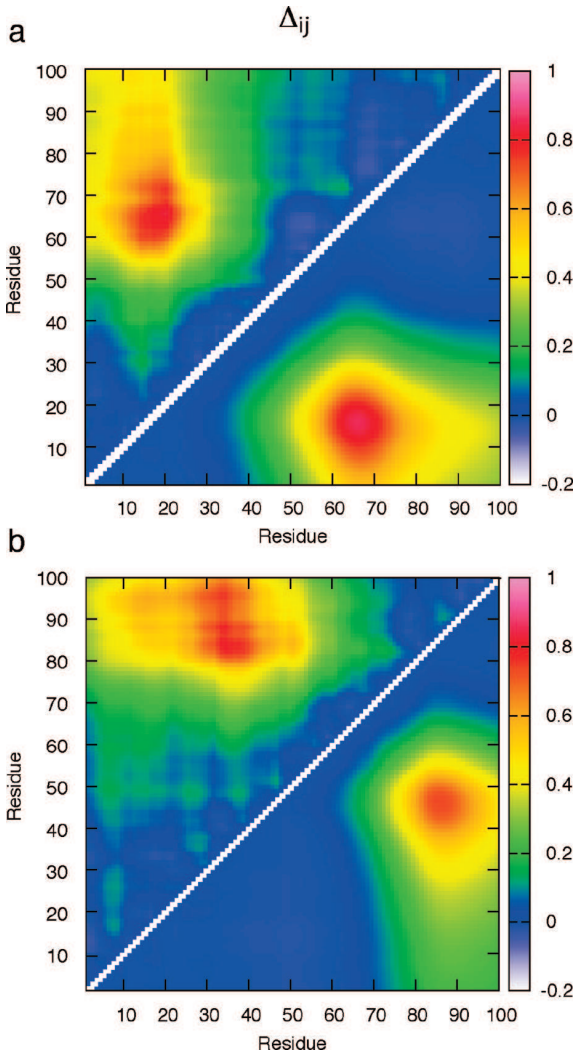

Figure 3. Contact maps showing chain proximity in the ensembles selected using ASTEROIDS on the basis of the data shown in Figure 2 (above the diagonal) in comparison with target ensembles (below the diagonal). In (a), the contact was between 11−20 and 61−70, while for (b), the contact was between 41−50 and 81−90. The scale for the data above the diagonal in each panel has been multiplied by a factor of 0.50 for ease of identification of the contact.

target ensemble. The previously noted tendency of PRE-based analysis to produce unrealistically compact ensembles of unfolded states, although present, was apparently less pronounced using the combined ASTEROIDS and *flexible-meccano* approach than in the case of restrained MD-based studies.[26−28] The exact origin of this observation is not clear and will require further comparative studies, but the improvement may be related to the explicit modeling of side-chain flexibility or to the fact that this approach uses the data to select representative ensembles rather than fitting the conformational sampling directly to the data.

We tested the ability of the combined ASTEROIDS and *flexible-meccano* approach to reproduce more than one contact. Clearly, the accuracy of this reproduction depends strongly on the number of paramagnetic probes and their specific distribution in the protein as well as the nature of the contacts (diffuse or well-defined). We performed an additional simulation, in this case for a protein containing 200 amino acids, where the target ensemble consisted of conformers with a contact between 11−20 and 61−70 or between 141−150 and 181−190. Simulated data from eight paramagnetic probes allowed ASTEROIDS to accurately and unambiguously find both contacts (Figure 5). The simulated target and fitted data from the eight sites are shown in Figure S1 in the Supporting Information.

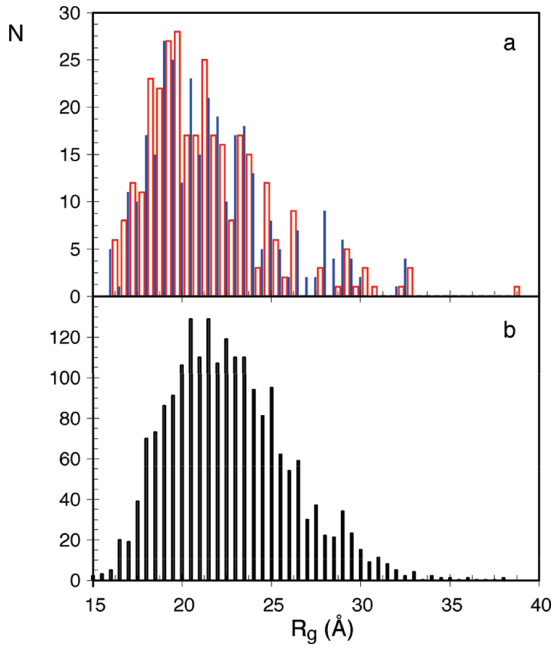

Figure 4. Ability of the ASTEROIDS approach to accurately reproduce the distribution of radii of gyration (R_g) in the selected ensembles. (a) Histogram showing the overall dimensions of the structures in ASTEROIDS ensembles selected on the basis of PREs shown in Figure 2b (contacts between 11−20 and 61−70). Blue: distribution of R_g in ensembles of size 80 (average R_g = 21.3 Å). Red: distribution of radii of gyration in ensembles of size 160 (average R_g = 21.7 Å). (b) Distribution of R_g for a set of 2000 structures from the target ensembles in which all of the structures contain a contact between 11−20 and 61−70 (average R_g = 22.6 Å).

Paramagnetic Relaxation Enhancement in Highly Disordered Systems: Experimental Data. In order to test the ensemble selection procedure further, we applied this approach to an experimental data set measured by Bertoncini et al.[23] for the intrinsically disordered protein α-synuclein. We employed these experimental data to determine how the use of an explicit flexible side-chain description of the spin label compares to using a fixed single position for each *flexible-meccano* conformer. In order to do this, we used ASTEROIDS to select ensembles based on the PRE data from cysteine mutants 18, 90, and 140 and then used these ensembles to predict the PREs measured for the spin label at position 76. It should be noted that this involved removing 25% of the available experimental data. The ensembles determined using a flexible side-chain description and a static side-chain description both fit the experimental data from the three "active" labels to within the experimental uncertainty, with the flexible side-chain model affording a slightly better fit (data not shown). More importantly, the reproduction of the "passive" data (i.e., the data not used in the ensemble selection) was systematically and significantly better when the flexible side-chain model was employed: the root-mean-square deviation (rmsd) for the flexible side-chain model was 0.17 ± 0.01, compared with an rmsd of 0.24 ± 0.02 for the static description. An example is shown in Figure 6, where the data reproductions of the PREs induced by the spin label at position 76 are compared for the two descriptions. This

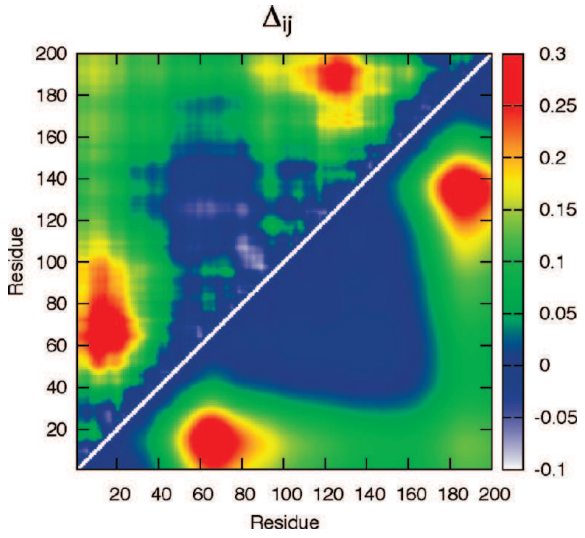

Figure 5. Contact map showing chain proximity in the presence of two contacts. Above the diagonal: contact map for an ASTEROIDS ensemble selected to reproduce simulated PRE data averaged over an ensemble in which each 200 amino acid conformer has a contact between 11−20 and 61−70 or between 141−150 and 181−190. In this case, eight PRE sites were simulated (sequence numbers 22, 44, 66, 88, 110, 132, 154, and 176). Below the diagonal: contact map for the target ensemble used to simulate the PRE data. The scale for the data above the diagonal has been multiplied by a factor of 0.66 for ease of identification of the contact.

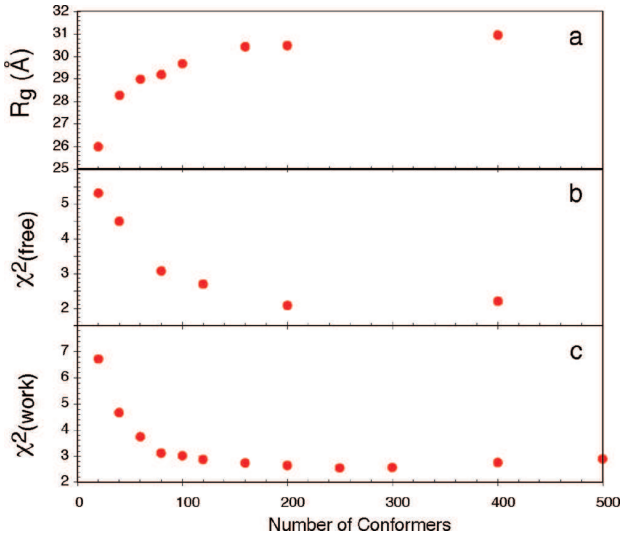

Figure 7. Ensemble characteristics as a function of selected ensemble size, targeting experimental PRE data measured in α-synuclein. (a) Average radius of gyration as a function of the number of structures in the selected ensemble. (b) χ^2 for the passive data as a function of the number of structures in the selected ensemble. The passive data in this case consists of the entire A76C data set. Only data from A18C, A90C, and A140C were used in the ensemble selection for the cross-validated reproduction of the "passive" data set. (c) χ^2 for the active data as a function of the number of structures in the selected ensemble.

example was chosen at random and is representative of the observed improvement. This result demonstrates the importance of incorporating local MTSL side-chain dynamics into the ensemble interpretation of the PREs, even for highly dynamic systems. These motions are predicted to occur on a relaxation-active time scale[45] and therefore require the use of the model-free or equivalent description that can explicitly account for the effect of local motions on the spectral density function. If fast motions of the spin label relative to the backbone are not included in the analysis, time-scale-dependent modulation of the observed relaxation interaction may be aliased into the effective intramolecular distance distribution.

The quality of the cross-validated data reproduction using the dynamic description allowed us to use this approach to probe the optimal number of structures required to describe the ensemble. This number depends on the complexity of the system (including the number of long-range contacts) as well as the number and position of the spin labels, but in this case, both the active and passive χ^2 values indicated that ensembles of ~200 structures were appropriate (Figure 7). This was supported by analysis of the effective radius of gyration, which rises until it reaches a plateau at approximately the same number of structures. Figure 8 shows the data reproduction when data from all four sites were included in the analysis; also shown is the resulting contact map comparing average interatomic distances in the ensemble with those from a control ensemble in which no selection on the basis of experimental data was made. In line with previous studies, a long-range contact between the C- and N-terminal domains was observed as well as a weaker contact between the so-called NAC region (residues 65−95) and the C-terminal domain.[22,23,35]

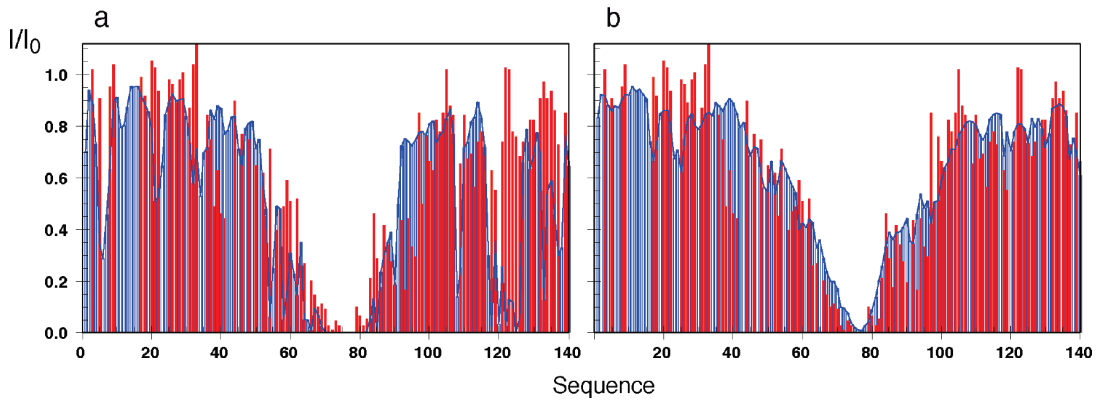

Figure 6. Cross-validation of "passive" α-synuclein PRE data. Only data from A18C, A90C, and A140C were used in the ensemble selection. (a) Example of the reproduction of the PRE data from the A76C site using the static position of the C^{β} atom as a representation of the average position of the spin label. (b) Example of the reproduction of the PRE data from the A76C site using the explicit MTSL side-chain dynamic averaging model described in the text. In both cases, the experimental PREs are shown in red and the calculated ratios in blue.

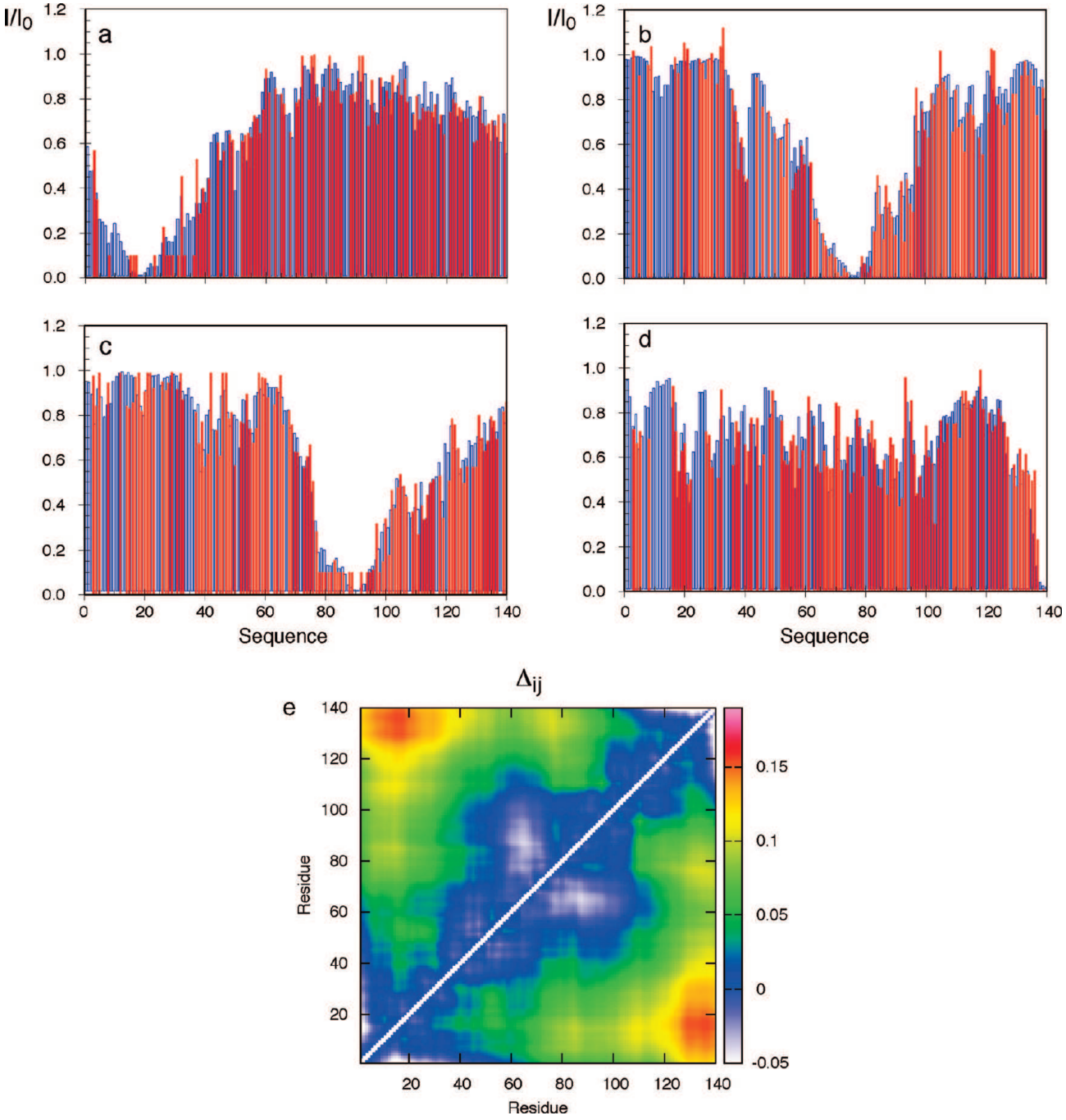

Figure 8. Reproduction of PRE data measured for α-synuclein. (a−d) Comparison of (red) experimental and (blue) ensemble-averaged data for an example calculation. (e) Resulting contact map showing the relative proximity of different parts of the chain.

Effects of Weak Long-Range Contacts on RDCs Measured in Highly Disordered Systems. In order to obtain a unified representation of the behavior of disordered proteins in solution, it is necessary to incorporate data from different sources that exhibit different structural and dynamic dependences. Here we investigate the effects of weak long-range contacts on the expected values of RDCs that are generally assumed to report mainly on local conformational propensities in disordered chains, and we propose appropriate guidelines for combining PREs and RDCs when using ensemble descriptions of flexible proteins.

The *flexible-meccano* approach was used to predict RDCs from 100 000-member ensembles of the 100 amino acid model sequence in the presence of weakly defined long-range contacts (Figure 9). The expected profiles when no specific contacts were present are also shown (Figure 9a). Figure 9b−g shows profiles of the expected $^{15}N-^{1}H^{N}$ ($^{1}D_{NH}$) and $^{13}C^{\alpha}-^{1}H^{\alpha}$ ($^{1}D_{C\alpha H\alpha}$) RDCs when a contact between two 20 amino acid strands (e.g., regions 1−20 and 81−100) was present. The effect of even such diffuse long-range contacts is surprisingly strong, resulting in significant quenching of the RDC values in regions between the two contact regions and some reinforcement of RDCs in the region of the contacting parts of the chain. Amino acids in all regions had essentially identical conformational sampling in all cases, but the RDCs were very different, indicating very clearly that

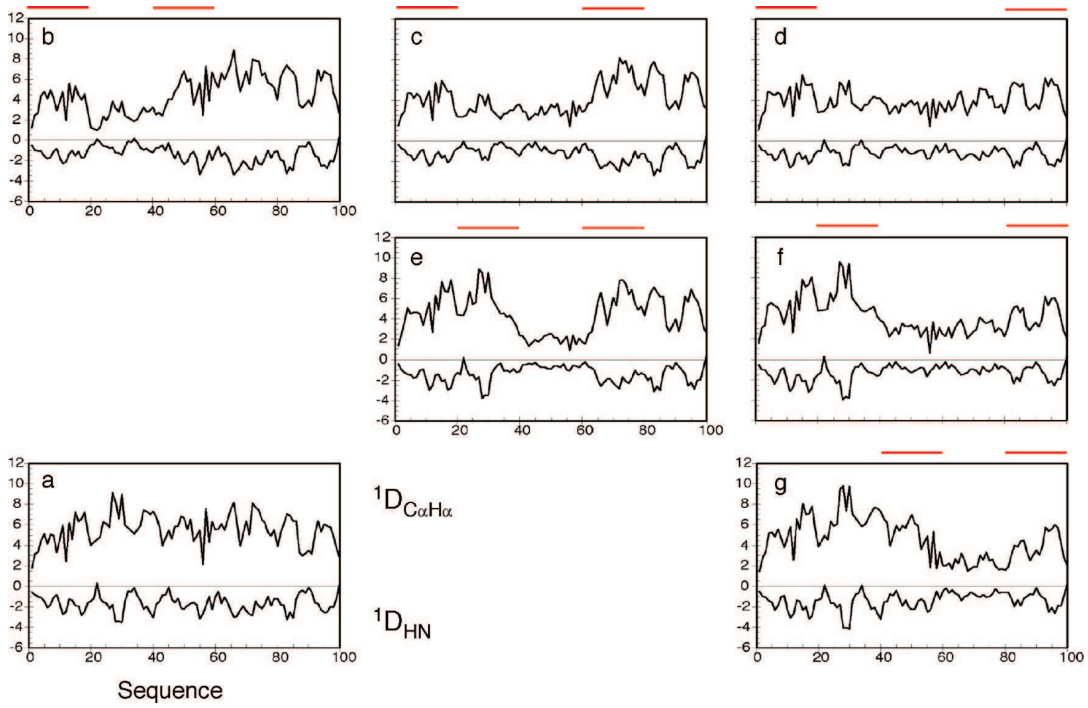

Figure 9. Simulation of $^1D_{NH}$ and $^1D_{C\alpha H\alpha}$ RDC profiles for a disordered protein with an arbitrary sequence in the presence of contacts between different sections of the chain. (a) Profile of couplings in the absence of specific contacts. The program PALES was used to calculate RDCs for each conformer; 100 000 conformers were used in this average and the ones shown in panels (b−g). (b−g) Profiles of couplings in the presence of contacts between regions i and j: (b) $i = 1-20$, $j = 41-60$; (c) $i = 1-20$, $j = 61-80$; (d) $i = 1-20$, $j = 81-100$; (e) $i = 21-40$, $j = 61-80$; (f) $i = 21-40$, $j = 81-100$; (g) $i = 41-60$, $j = 81-100$. The two continuous red bars above each plot indicate the positions of the contacting regions.

caution needs to be exercised when interpreting RDCs uniquely in terms of local structure if long-range contacts are also present. This would potentially lead to significant error in the cases shown in Figure 9.

In order to further clarify the origin of these effects, the same analysis was carried out for a homopolymer (polyvaline), resulting in the expected bell-shaped curve for the ensembles without contact-specific selection (Figure 10a) and clear modifications occurring for the ensembles with specific contacts (Figure 10b−g). The effect of diffuse long-range contacts is apparently to superpose a more complex baseline upon the local structure of the expected RDCs. This baseline has peaks in the interacting regions and a trough in the intervening region. We believe that the effect has a similar origin as that found in the presence of helical elements in disordered chains, where $^1D_{NH}$ values become positive as a result of the effective average alignment of the $^{15}N-^1H^N$ bond vectors with the average chain direction and thereby the magnetic field.[60] The same effect may occur here, although in this case, the helix has a very long period in terms of amino acids and therefore would create a very broad inverted curve relative to the bell-shaped curve, whose shallowness depends on the distance between the interacting segments, as observed from the numerical simulation.

Parametrization of the Effect of Long-Range Contacts on RDCs in Disordered Systems. It has previously been shown that RDCs from unfolded chains with no specific interacting regions can be expressed in terms of the product of a generic baseline, b_{ml}, and RDCs derived from sampling of conformational space that can be defined using short local alignment windows (LAWs):[59,43]

$$D_{ml} = |b_{ml}| D_{ml}^{LAW} \quad (5)$$

where m and l represent the pair of nuclei (e.g., N and H^N). In Figure 10, the red curves were obtained using the parametrization of a generic baseline expression that reproduces the numerically predicted baselines shown for the polyvaline chain (see Methods for the full expression). This can be described as a combination of the baseline expression for no specific contacts (a hyperbolic cosine function introduced previously[43]) with Gaussian curves between the contact points. Importantly, the curves depend only on the position of the contacts and the length of the chain.

This expression can then be combined with RDCs predicted using LAWs accounting for short-range conformational behavior. This is illustrated in Figure 11, where the LAW-derived profile (Figure 11b), which was calculated using 200 structures, is combined with the baseline predicted for long-range contacts between segments 41−60 and 81−100 (Figure 11a). The prediction agrees essentially identically with the explicit simulations calculated using 100 000 conformers containing the required contact (Figure 11c). In the case of more than one contact (as shown in Figure 5, for example), the baseline effects are combined as shown in eq 11 and can again be shown to accurately reproduce the effects simulated from explicit averages over 100 000 conformers containing these contacts (see Figure S2 in the Supporting Information).

Simultaneous Analysis of PRE-Derived Long-Range Contacts and RDC-Derived Local Information. The above results show that it is possible in principle to combine PRE-derived long-range information with RDC-derived local information while accounting for possibly significant long-range effects on RDCs and preserving a relatively small number of structures. This latter point is of particular importance when using ensemble selection approaches. In order to test this possibility further, we analyzed

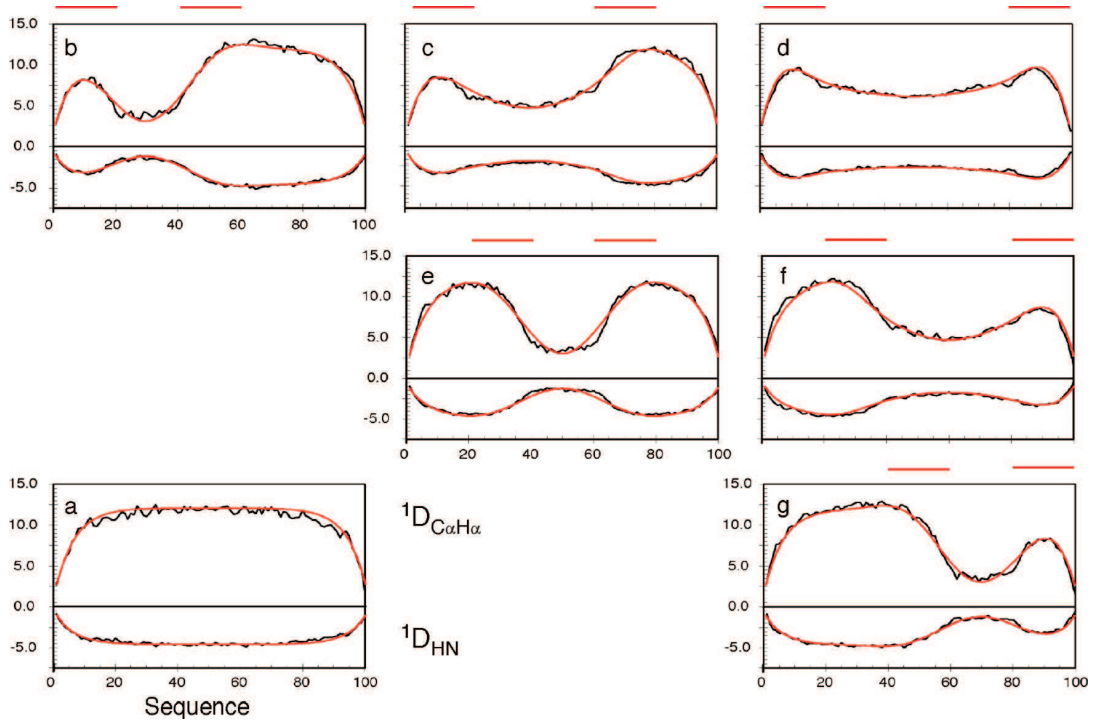

Figure 10. Simulation of RDC profiles for a homopolymer (polyvaline) in the presence of contacts between different sections of the chain. (a) Profile of calculated couplings in the absence of specific contacts. The program PALES was used to calculate RDCs from each conformer; 100 000 conformers were used in this average and the ones shown in panels (b−g). (b−g) Profiles of couplings in the presence of contacts between regions i and j: (b) $i = 1-20$, $j = 41-60$; (c) $i = 1-20$, $j = 61-80$; (d) $i = 1-20$, $j = 81-100$; (e) $i = 21-40$, $j = 61-80$; (f) $i = 21-40$, $j = 81-100$; (g) $i = 41-60$, $j = 81-100$. The two continuous bars above each plot indicate the positions of the contacting regions. The red curves were computed using eq 11 with the contact positioned in the center of each region.

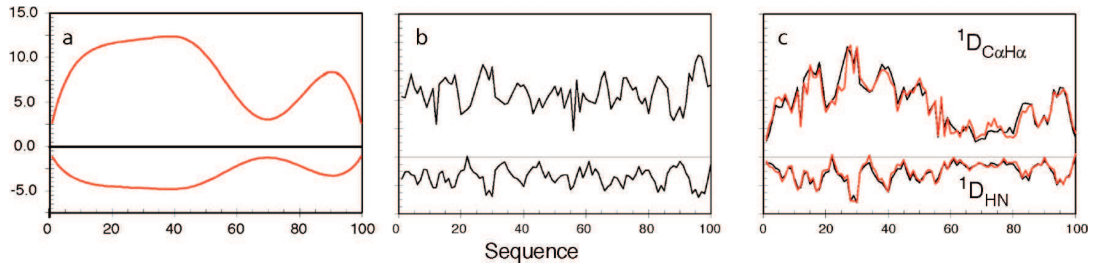

Figure 11. Example of the combination of analytically calculated baselines and RDCs averaged using the local alignment window (LAW) approach. (a) Baseline contribution calculated analytically using eq 11 for contacts between the regions centered on residues 50 and 90. (b) RDCs calculated using the previously proposed LAW approach with windows 15 amino acids in length; each RDC was averaged over 200 structures. (c) Combination of the baseline from (a) and the local RDCs from (b) (red curves) compared to the RDCs averaged over 100 000 full-length conformers in which each structure has a contact between 41−60 and 81−100 (black curves).

the ensembles presented in Figure 3, where the target contacts were between positions 11−20 and 61−70 and between positions 41−50 and 81−90. The contact matrices were analyzed to find the maximum of the difference between the PRE-derived ensemble and the reference ensemble containing no specific contacts (see Methods). The results are shown in Figure 12. In Figure 12a, the red and blue curves indicate the RDC baselines derived using this approach (calculated using eq 11), and the black curve shows the $^1D_{NH}$ RDCs calculated using the LAW approach. In Figure 12b,c, the combination of the baseline and the locally calculated RDCs is compared to RDCs calculated explicitly from 100 000 conformers, all of which fulfill the contact criterion. The good agreement demonstrates that one can combine PREs and RDCs in a meaningful way for the ensemble description of disordered proteins using experimental data.

Combining Experimental PREs and RDCs in α-Synuclein Validates RDC Baseline Analysis. Finally, we applied this analysis to the contact matrix determined on the basis of experimental PRE data from α-synuclein (shown in Figure 8e). Experimentally measured RDCs are shown in Figure 13a and compared to RDCs calculated from an explicit representation of full-length α-synuclein. The RDC baseline derived from analysis of the contact matrix is shown in Figure 13b, superimposed on the RDCs calculated using the LAW approach. The two curves were combined using eq 5, and the result is compared to the experimental data (after appropriate scaling) in Figure 13c. The RDC profile reproduces the experimental data significantly better than the ensemble derived in the absence of specific contacts (rmsd of 0.52 Hz compared with 0.78 Hz). This study therefore not only validates the predicted effects on

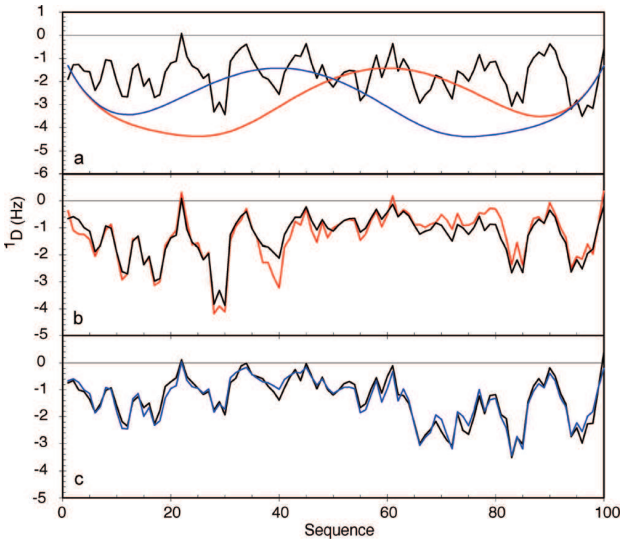

Figure 12. Example of a combined analysis of PREs and RDCs in the context of simulated data. PREs were used to determine long-range contacts. RDC profiles were calculated using baselines determined on the basis of PRE analysis and LAWs. Contacts were identified from distance matrices as described in the text. The reproduction of the PREs and the resulting distance matrix from this simulation are shown in Figures 2 and 3. (a) Black curve: LAW-averaged RDCs. Blue curve: RDC baseline extracted from the contact matrix shown in Figure 3a (contact between 11−20 and 61−70). Red curve: RDC baseline extracted from the contact matrix shown in Figure 3b (contact between 41−50 and 81−90). (b) Black curve: RDCs calculated from an explicit ensemble calculation using 100 000 conformers containing a contact between 41−50 and 81−90. Red curve: the combination of the LAW curve and red baseline curve shown in (a) (contact between regions 41−50 and 81−90). (c) Black curve: RDCs calculated from an explicit ensemble calculation using 100 000 conformers containing a contact between 11−20 and 61−70. Red curve: combination of the LAW curve and blue baseline curve shown in (a) (contact between regions 11−20 and 61−70).

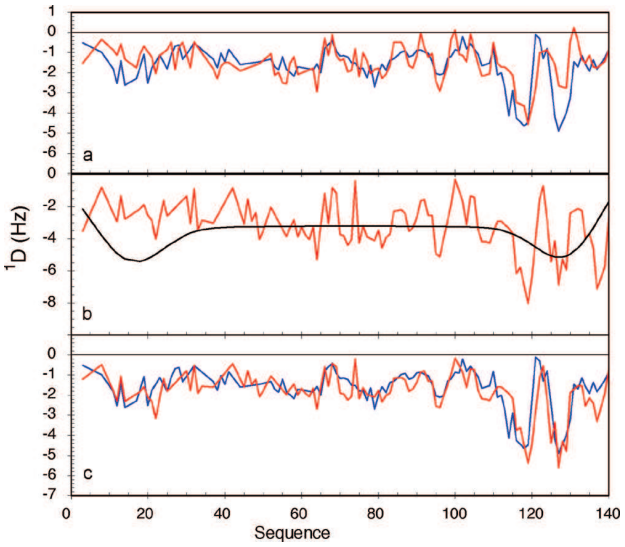

Figure 13. Example of a combined analysis of PREs and RDCs in the context of experimental data: comparison of experimental $^1D_{NH}$ RDCs measured from α-synuclein aligned in PEG-hexanol with values obtained using the combination of LAW and baseline prediction from PRE analysis. (a) Comparison of experimental $^1D_{NH}$ RDCs (blue) with couplings calculated using a standard *flexible-meccano* prediction (red). The rmsd between the two distributions was 0.78 Hz. (b) LAW-predicted RDCs (red) and effective baseline derived from the contact map shown in Figure 8e using eq 11 (black). (c) Combination of the curves shown in (b) (red) compared to the experimental $^1D_{NH}$ RDCs (blue). The rmsd in this case was 0.52 Hz.

RDC profiles due to long-range transient contacts in unfolded systems but also demonstrates that PREs and RDCs can be usefully combined in an experimental context. This provides further support for previously published observations that RDCs have been correctly reproduced only in the presence of long-range contacts.[35]

Conclusions

In order to understand the conformational behavior of IDPs, a molecular representation of the partially folded state is required. Because of the very large number of degrees of conformational freedom available to such a disordered system, this representation should be based on extensive sets of experimental data. Each experimental parameter is sensitive to different aspects of the structural and dynamic behavior of the disordered state and requires specific consideration of the relevant averaging properties of the physical interaction. In this study, we have taken another step toward the development of a unified molecular representation of the disordered state by combining complementary data sets with novel analytical tools designed to exploit the specific conformational sensitivity of the different experimental parameters.

Having recently demonstrated that multiple RDCs can be combined with an efficient ensemble selection algorithm (ASTEROIDS) to define local conformational sampling directly from the experimental data, we have extended the approach to incorporate the possible presence of long-range contacts. We have demonstrated the use of ASTEROIDS to analyze PREs and faithfully reproduce intramolecular proximity even in the presence of highly diffuse, ill-defined contacts that give rise to broad PRE profiles. We have also demonstrated that the combination of numerical and analytical modeling of spin-label mobility significantly improves the reproduction of the experimental data. The effects of long-range contacts on RDCs have been shown to produce severe distortion of RDC profiles predicted on the basis of local sampling alone. We have demonstrated that this distortion can be generally parametrized and combined with RDC prediction based on local sampling alone to provide an efficient and reliable tool for interpreting RDCs in flexible chains containing preferred long-range contacts.

We thus have shown that it is possible to combine NMR data that exhibit very different averaging properties and structural dependences in a meaningful way, providing the perspective of characterizing the essential local and long-range conformational characteristics of unfolded proteins using PREs and RDCs. In the example we provided, the reproduction of experimental RDCs from the protein α-synuclein was significantly improved when baseline effects derived from the PRE analysis were introduced into the analysis, demonstrating the feasibility of combining these experimental parameters into an informative ensemble description.

Methods

Experimental Data. Details of experimental measurements of RDCs and PREs have been published elsewhere.[23,33]

PRE Calculations with *Flexible-Meccano*. Sterically allowed MTSL side-chain conformations were sampled using previously published rotameric distributions[68] and built explicitly for each spin-label site of each *flexible-meccano* backbone; 600 side-chain conformers were calculated, and the sterically allowed conformers were retained. Relaxation effects were averaged over these conformers as described in Theoretical Aspects.

Definition of Contacts. We considered a contact to be present between two different parts of the polypeptide chain if the C^{β} of an amino acid in one contiguous strand (e.g., residues 11−20) was located less than 15 Å from any C^{β} in another contiguous strand (e.g., residues 51−60).

Contact Matrices. Average distances between sites were represented in terms of the metric Δ_{ij}, defined as

$$\Delta_{ij} = \log(\langle d_{ij}\rangle/\langle d_{ij}^0\rangle) \tag{6}$$

where d_{ij} is the distance between sites i and j in any given structure of the ASTEROIDS ensemble and d_{ij}^0 is the distance between sites i and j in any given structure of the reference ensemble (with no specific selection). This metric was used to highlight a higher propensity to form contacts than in a molecule that has no specific contacts. It should be noted that this representation of average interatomic distances naturally (and artificially) enhances contacts that are further apart in the chain, so the observed contacts are "smeared" away from the diagonal.

Contact matrices were analyzed to determine $I_{ij}^{\max}$, the maximum of the difference between the PRE-derived ensemble and the reference ensemble containing no specific contacts:

$$I_{ij}^{\max} = \max_{i,j\in[1,n]}(\Delta_{ij}) \tag{7}$$

The matrix was divided into segments of 5 × 5 amino acids and searched for the highest-populated segment fulfilling the following criterion:

$$0.9I_{ij}^{\max} \le \Delta_{ij} \le I_{ij}^{\max} \tag{8}$$

This approach identified the highest-populated contacting region. The center of this region was then used to calculate the baseline effects on the RDC profile using eq 11.

RDC Calculations with *Flexible-Meccano* Using a Global Alignment Tensor. Simulated RDCs were calculated using the program *flexible-meccano* interfaced to PALES.[69] Profiles of RDCs in the presence of long-range order were simulated by retaining only conformers for which the desired contact was present.

RDC Calculations with *Flexible-Meccano* Using a Local Alignment Window. For calculations using a LAW, the RDC for the central amino acid of the local 15 amino acid segment was calculated for each individual structure.[43] For the terminal amino acids, seven alanines were added to the N- or C-terminus during the building of the protein to ensure that a 15 amino acid segment was always present. The resulting RDC profile along the primary sequence was calculated by averaging each value over the whole ensemble and multiplying by the corresponding scaled absolute value of the effective baseline given in eq 11. RDCs calculated using full-length descriptions of the protein were averaged over all conformers as previously described.[34]

ASTEROIDS Ensemble Selection. ASTEROIDS uses a previously described genetic algorithm to build a representative ensemble of structures of fixed size N from a large database. The algorithm selects an ensemble of N structures by comparing with experimental data using the following fitness function:

$$\chi^2_{\text{ASTEROIDS}} = \sum_k (\Delta I^k_{\text{calcd}} - \Delta I^k_{\text{exptl}})^2 \tag{9}$$

where

$$\Delta I^k_{\text{calcd}} = \frac{\Gamma^k_{2,\text{red}}\exp(-\Gamma^k_{2,\text{para}}t_m)}{\Gamma^k_{2,\text{red}} + \Gamma^k_{2,\text{para}}} \tag{10}$$

in which $\Gamma_{2,\text{para}}$ is the paramagnetic component of the measured relaxation rate given in eq 4, $\Gamma_{2,\text{red}}$ is the intrinsic transverse relaxation rate of the observed proton spin, and t_m is the mixing time, for which a value of 10 ms was used. The final ensemble is obtained from generations of ensembles that undergo evolution and selection using this fitness function. Each generation comprises 100 different ensembles of size N. Remaining parameters are treated as previously described.

Parametrization of a Generic RDC Baseline Expression for Transiently Contacting Chains. A generic RDC baseline expression for transiently contacting chains can be obtained by combining the baseline expression for no specific contacts (a hyperbolic cosine function introduced previously[43]) with a Gaussian curve between the contact points and then correcting this with Gaussian curves in the vicinity of the contacting points. Importantly, the Gaussian curves depend only on the position of the contacts and the length of the chain. This results in the following analytical expression for the baseline RDC, D_{ij}^{BL}:

$$\begin{aligned} D_{ij}^{\text{BL}} = \{2b(L)\cosh[-a(L)(m - m_0)] - \\ c(L)\}\Big(1 - \sum_i \{G_i e^{-(m-n_{0,i})^2/2\sigma_i^2} + \\ H_i\big[(D_i + S_i)e^{-(m-n_{1,i}+S_i/2)^2/2\delta^2} + \\ (D_i - S_i)e^{-(m-n_{2,i}+S_i/2)^2/2\delta^2}\big]\}\Big) \end{aligned} \tag{11}$$

where L is the length of the chain, the contact occurs between positions n_1 and n_2, and the sum includes all of the independent contacts i. Other parameters are defined as follows: $m_0 = (L + 1)/2$, $n_0 = (n_1 + n_2)/2$, $D = |n_1 - n_2|$, and $S = n_0 - m_0$. The parametrizations of a, b, c, G, H, σ, and δ are given in the Supporting Information.

Acknowledgment. L.S. received a grant from the French Ministry of Education. This work was supported by the French Research Ministry through ANR Protein Motion PCV07_194985 PCVI 0013 and the Deutsche Forschungsgemeinschaft Heisenberg Scholarship Z.W. 71/2-1 and 3-1 (to M.Z). M.R.J. benefited from a long-term EMBO fellowship and Lundbeckfonden support.

Supporting Information Available: Figure S1 showing a reproduction of simulated sample PRE data for ensembles containing two specific contacts (produced using the ensemble selection algorithm ASTEROIDS) and the associated baseline effects; Figure S2 showing a comparison between RDCs calculated by ensemble averaging and the baseline contribution calculated using eq 11; Figure S3 showing RDCs measured in A76C cysteine mutant and wild-type α-synuclein; Figure S4 showing calculated and experimental ^3J scalar couplings from α-synuclein; and parametrization of a generic RDC baseline expression. This material is available free of charge via the Internet at http://pubs.acs.org.

JA101645G

(69) Zweckstetter, M.; Bax, A. *J. Am. Chem. Soc.* **2000**, *122*, 3791.

BIBLIOGRAPHY

[1] C. A. Ronan. Histoire mondiale des sciences. *Points Seuil*, 1988. (cited p. 3.)

[2] J. Yon-Kahn. Histoire de la science des protéines. *EDP Sciences*, 2006. (cited p. 3 and 4.)

[3] A. Brack and F. Raulin. L'évolution chimique et les origines de la vie. *Masson*, 1991. (cited p. 3.)

[4] N. A. Campbell and J. B. Reece. Biologie. *De Boeck*, 2004. (cited p. 3, 4 and 146.)

[5] P. Anderson. More is different. *Science*, 177(4047):393–396, 1972. (cited p. 4.)

[6] J. M. Berg, J. L. Tymoczko, and L. Stryer. Biochemistry. *W.H. Freeman*, 2002. (cited p. 4 and 146.)

[7] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. Molecular biology of the cell. *Garland Science*, 2008.

[8] H. Lodish, A. Berk, C. A. Kaiser, M. Krieger, M. P. Scott, and A. Bretscher. Molecular cell biology. *W.H. Freeman*, 2008. (cited p. 4, 146 and 261.)

[9] A. Mirsky and L. Pauling. On the structure of native, denatured, and coagulated proteins. *Proc Natl Acad Sci USA*, 22(7):439, 1936. (cited p. 4.)

[10] L. Pauling, R. B. Corey, and H. R. Branson. The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proc Natl Acad Sci USA*, 37(4):205–11, 1951. (cited p. 4.)

[11] L. Pauling and R. B. Corey. Configurations of polypeptide chains with favored orientations around single bonds: Two new pleated sheets. *Proc Natl Acad Sci USA*, 37(11):729–40, 1951. (cited p. 4.)

[12] J. Watson and F. H. Crick. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–8, 1953. (cited p. 4.)

[13] A. Abragam. The principles of nuclear magnetism. *Oxford University Press*, 1989. (cited p. 5, 6, 11, 13, 14, 17, 18, 29 and 30.)

[14] C. Cohen-Tannoudji, B. Diu, and F. Laloë. Mécanique quantique i. *Hermann*, 1973. (cited p. 11, 12, 13, 14, 15, 35 and 42.)

[15] M. H. Levitt. Spin dynamics: basics of nuclear magnetic resonance. *Wiley*, 2008. (cited p. 5, 11, 12, 14, 18, 19, 20, 21, 29, 30, 31, 32, 36, 37, 38 and 193.)

[16] R. R. Ernst, G. Bodenhausen, and A. Wokaun. Principles of nuclear magnetic resonance in one and two dimensions. *Oxford University Press*, 1990. (cited p. 6, 13, 18, 19, 20 and 37.)

[17] M. Goldman. Quantum description of high-resolution nmr in liquids. *Oxford University Press*, 1991. (cited p. 6, 11, 13, 14, 15, 17, 18, 19, 29, 35 and 36.)

[18] A. G. Palmer. Nmr characterization of the dynamics of biomacromolecules. *Chem Rev*, 104(8):3623–40, 2004. (cited p. 6, 19, 24 and 27.)

[19] M. Blackledge. Recent progress in the study of biomolecular structure and dynamics in solution from residual dipolar couplings. *Prog. NMR Spectrosc.*, 46:23–61, 2005. (cited p. 6, 30, 32, 39, 47 and 65.)

[20] N. Bloembergen, E. Purcell, and R. Pound. Relaxation effects in nuclear magnetic resonance absorption. *Phys Rev*, 73(7):679–712, 1948. (cited p. 13.)

[21] D. Korzhnev, M. Billeter, A. Arseniev, and V. Orekhov. Nmr studies of brownian tumbling and internal motions in proteins. *Prog. NMR Spectrosc.*, 38(3):197, 2001. (cited p. 15, 22, 23, 24 and 149.)

[22] L. Werbelow. Nmr dynamic frequency shifts and the quadrupolar interaction. *J. Chem. Phys.*, 70:5381–5383, 1979. (cited p. 17.)

[23] R. Brüschweiler. Cross-correlation-induced j coupling. *Chem Phys Lett*, 257 (1-2):119–122, 1996. (cited p. 17.)

[24] M. Fischer, A. Majumdar, and E. R. P. Zuiderweg. Protein nmr relaxation: theory, applications and outlook. *Prog. NMR Spectrosc.*, 33(3-4):207–272, 1998. (cited p. 17, 18, 19, 22, 24 and 149.)

[25] R. Atkinson and B. Kieffer. The role of protein motions in molecular recognition: insights from heteronuclear nmr relaxation measurements. *Prog. NMR Spectrosc.*, 44(3-4):141–187.

[26] V. A. Jarymowycz and M. J. Stone. Fast time scale dynamics of protein backbones: Nmr relaxation methods, applications, and functional consequences. *Chem Rev*, 106(5):1624–71, 2006. (cited p. 18, 19, 22, 23, 24 and 149.)

[27] J. Schotland and J. Leigh. Exact solutions of the bloch equations with n-site chemical exchange. *J Magn Res*, 51(1):48–55, 1983. (cited p. 19.)

[28] A. Bain. Chemical exchange in nmr. *Prog. NMR Spectrosc.*, 43(3-4):63–103, 2003. (cited p. 19 and 27.)

[29] A. G. Palmer, C. D. Kroenke, and J. P. Loria. Nuclear magnetic resonance methods for quantifying microsecond-to-millisecond motions in biological macromolecules. *Meth Enzymol*, 339:204–38, 2001. (cited p. 20, 27 and 168.)

[30] V. Daragan and K. Mayo. Using the model free approach to analyze nmr relaxation data in cases of anisotropic molecular diffusion. *J. Phys. Chem. B*, 103(32):6829–6834, 1999. (cited p. 22.)

[31] D. Woessner. Nuclear spin relaxation in ellipsoids undergoing rotational brownian motion. *J. Chem. Phys.*, 37:647, 1962. (cited p. 22.)

[32] J. G. de la Torre, M. Huertas, and B. Carrasco. Calculation of hydrodynamic properties of globular proteins from their atomic-level structure. *Biophys J*, 78 (2):719–730, 2000. (cited p. 23.)

[33] P. Bernado, J. G. de la Torre, and M. Pons. Interpretation of 15n nmr relaxation data of globular proteins using hydrodynamic calculations with hydronmr. *J Biomol NMR*, 23(2):139–50, 2002. (cited p. 23.)

[34] D. Fushman, R. Varadan, M. Assfalg, and O. Walker. Determining domain orientation in macromolecules by using spin-relaxation and residual dipolar coupling measurements. *Prog. NMR Spectrosc.*, 44(3):189–214, 2004. (cited p. 23.)

[35] D. Fushman, R. Xu, and D. Cowburn. Direct determination of changes of interdomain orientation on ligation: Use of the orientational dependence of 15n nmr relaxation in abl sh(32). *Biochemistry*, 38(32):10225–10230, 1999.

[36] R. Ghose, D. Fushman, and D. Cowburn. Determination of the rotational diffusion tensor of macromolecules in solution from nmr relaxation data with a combination of exact and approximate methods–application to the determination of interdomain orientation in multidomain proteins. *J Magn Res*, 149(2):204–217, 2001. (cited p. 23.)

[37] V. Daragan and K. Mayo. Motional model analyses of protein and peptide dynamics using c-13 and n-15 nmr relaxation. *Prog. NMR Spectrosc.*, 31(1): 63–105, 1997. (cited p. 23.)

[38] R. London and J. Avitabile. Calculated carbon-13 nmr relaxation parameters for a restricted internal diffusion model. application to methionine relaxation in dihydrofolate reductase. *J Am Chem Soc*, 100(23):7159–7165, 1978. (cited p. 23.)

[39] T. Bull, J. Norne, P. Reimarsson, and B. Lindman. Nuclear magnetic resonance studies of chloride binding to proteins. *J Am Chem Soc*, 100(15):4643–4647, 1978. (cited p. 23.)

[40] G. M. Clore, P. C. Driscoll, P. T. Wingfield, and A. M. Gronenborn. Analysis of the backbone dynamics of interleukin-1 beta using two-dimensional inverse detected heteronuclear 15n-1h nmr spectroscopy. *Biochemistry*, 29(32):7387–401, 1990. (cited p. 23 and 24.)

[41] T. Bremi and R. Brüschweiler. Locally anisotropic internal polypeptide backbone dynamics by nmr relaxation. *J Am Chem Soc*, 119(28):6672–6673, 1997. (cited p. 23, 53 and 97.)

[42] S. F. Lienin, T. Bremi, B. Brutscher, R. Brüschweiler, and R. R. Ernst. Anisotropic intramolecular backbone dynamics of ubiquitin characterized

by nmr relaxation and md computer simulation. *J Am Chem Soc*, 120(38): 9870–9879, 1998. (cited p. 23.)

[43] G. M. Clore, A. Szabo, A. Bax, L. E. Kay, P. Driscoll, and A. M. Gronenborn. Deviations from the simple two-parameter model-free approach to the interpretation of nitrogen-15 nuclear magnetic relaxation of proteins. *J Am Chem Soc*, 112(12):4989–4991, 1990. (cited p. 24 and 25.)

[44] G. Lipari and A. Szabo. Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. theory and range of validity. *J Am Chem Soc*, 104(17):4546–4559, 1982. (cited p. 24.)

[45] G. Lipari and A. Szabo. Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 2. analysis of experimental results. *J Am Chem Soc*, 104(17):4559–4570, 1982. (cited p. 24.)

[46] B. Halle. The physical basis of model-free analysis of nmr relaxation data from proteins and complex fluids. *J. Chem. Phys.*, 131(22):224507, 2009. (cited p. 24, 38 and 224.)

[47] P.-G. de Gennes and J. Prost. The physics of liquid crystals. *Oxford University Press*, 1995. (cited p. 29 and 33.)

[48] J. H. Prestegard, H. M. Al-Hashimi, and J. R. Tolman. Nmr structures of biomolecules using field oriented media and residual dipolar couplings. *Q Rev Biophys*, 33(4):371–424, 2000. (cited p. 30, 32 and 65.)

[49] A. Saupe and G. Englert. High-resolution nuclear magnetic resonance spectra of orientated molecules. *Phys Rev Lett*, 11(10):462–464, 1963. (cited p. 30.)

[50] J. Lohman and C. MacLean. Alignment effects on high resolution nmr spectra, induced by the magnetic field. *Chem Phys*, 35(3):269–274, 1978. (cited p. 30.)

[51] M. Lisicki, P. Mishra, and A. A. Bothner-By. Solution conformation of a porphyrin-quinone cage molecule determined by dipolar magnetic field effects in ultra-high-field nmr. *J. Phys. Chem.*, 92(12):3400–3403, 1988. (cited p. 30.)

[52] J. R. Tolman, J. M. Flanagan, M. A. Kennedy, and J. H. Prestegard. Nuclear magnetic dipole interactions in field-oriented proteins: information for structure determination in solution. *Proc Natl Acad Sci USA*, 92(20):9279–83, 1995. (cited p. 30 and 32.)

[53] N. Tjandra and A. Bax. Direct measurement of distances and angles in biomolecules by nmr in a dilute liquid crystalline medium. *Science*, 278(5340): 1111–4, 1997. (cited p. 30 and 33.)

[54] S. Elavarasi, A. Kumari, and K. Dorai. Using the chemical shift anisotropy tensor of carbonyl backbone nuclei as a probe of secondary structure in proteins. *J. Phys. Chem. A*, 114(18):5830–5837, 2010. (cited p. 31.)

[55] R. Lipsitz and N. Tjandra. Carbonyl csa restraints from solution nmr for protein structure refinement. *J Am Chem Soc*, 123(44):11065–11066, 2001.

[56] E. Czinki, A. Császár, G. Magyarfalvi, P. Schreiner, and W. Allen. Secondary structures of peptides and proteins via nmr chemical-shielding anisotropy (csa) parameters. *J Am Chem Soc*, 129(6):1568–1577, 2007. (cited p. 31.)

[57] L. Yao, A. Grishaev, G. Cornilescu, and A. Bax. Site-specific backbone amide 15n chemical shift anisotropy tensors in a small protein from liquid crystal and cross-correlated relaxation measurements. *J Am Chem Soc*, 132(12):4295–4309, 2010. (cited p. 31.)

[58] I. Bertini, C. Luchinat, and G. Parigi. Magnetic susceptibility in paramagnetic nmr. *Prog. NMR Spectrosc.*, 40(3):249, 2002. (cited p. 32 and 33.)

[59] H. Kung, K. Wang, I. Goljer, and P. Bolton. Magnetic alignment of duplex and quadruplex dnas. *J Magn Res Ser B*, 109(3):323, 1995. (cited p. 32.)

[60] N. Tjandra, J. Omichinski, A. Gronenborn, G. Clore, and A. Bax. Use of dipolar 1h–15n and 1h–13c couplings in the structure determination of magnetically oriented macromolecules in solution. *Nat Struct Mol Biol*, 4(9): 732–738, 1997. (cited p. 32.)

[61] J. Feeney, B. Birdsall, A. Bradbury, R. Biekofsky, and P. Bayley. Calmodulin tagging provides a general method of using lanthanide induced magnetic field orientation to observe residual dipolar couplings in proteins in solution. *J Biomol NMR*, 21(1):41–48, 2001. (cited p. 32.)

[62] J. Wöhnert, K. Franz, M. Nitz, B. Imperiali, and H. Schwalbe. Protein alignment by a coexpressed lanthanide-binding tag for the measurement of residual dipolar couplings. *J Am Chem Soc*, 125(44):13338–13339, 2003. (cited p. 32.)

[63] C. Sanders and J. P. Schwonek. Characterization of magnetically orientable bilayers in mixtures of dihexanoylphosphatidylcholine and dimyristoylphosphatidylcholine by solid-state nmr. *Biochemistry*, 31:8898–8905, 1992. (cited p. 33.)

[64] C. Sanders, B. Hare, K. Howard, and J. H. Prestegard. Magnetically-oriented phospholipid micelles as a tool for the study of membrane-associated molecules. *Prog. NMR Spectrosc.*, 26:421–444, 1994. (cited p. 33.)

[65] C. Sanders and R. S. Prosser. Bicelles: a model membrane system for all seasons? *Structure*, 6(10):1227–34, 1998. (cited p. 33.)

[66] C. Sanders, J. E. Schaff, and J. H. Prestegard. Orientational behavior of phosphatidylcholine bilayers in the presence of aromatic amphiphiles and a magnetic field. *Biophys J*, 64(4):1069–80, 1993. (cited p. 33.)

[67] S. Cavagnero, H. Dyson, and P. E. Wright. Improved low ph bicelle system for orienting macromolecules over a wide temperature range. *J Biomol NMR*, 13(4):387–391, 1999. (cited p. 34.)

[68] M. Ottiger and A. Bax. Bicelle-based liquid crystals for nmr-measurement of dipolar couplings at acidic and basic ph values. *J Biomol NMR*, 13(2):187–191, 1999. (cited p. 34, 80, 81, 109 and 147.)

[69] M. Ottiger and A. Bax. Characterization of magnetically oriented phospholipid micelles for measurement of dipolar couplings in macromolecules. *J Biomol NMR*, 12(3):361–372, 1998. (cited p. 34.)

[70] H. Wang, M. Eberstadt, E. Olejniczak, R. Meadows, and S. Fesik. A liquid crystalline medium for measuring residual dipolar couplings over a wide range of temperatures. *J Biomol NMR*, 12(3):443–446, 1998. (cited p. 34.)

[71] J. A. Losonczi and J. H. Prestegard. Improved dilute bicelle solutions for high-resolution nmr of biological macromolecules. *J Biomol NMR*, 12(3):447–51, 1998. (cited p. 34 and 147.)

[72] B. Ramirez and A. Bax. Modulation of the alignment tensor of macromolecules dissolved in a dilute liquid crystalline medium. *J Am Chem Soc*, 120(35):9106–9107, 1998. (cited p. 34 and 147.)

[73] R. Prosser, S. Hunt, J. DiNatale, and R. Vold. Magnetically aligned membrane model systems with positive order parameter: switching the sign of szz with paramagnetic ions. *J Am Chem Soc*, 118(1):269–270, 1996. (cited p. 34.)

[74] G. M. Clore, M. Starich, and A. M. Gronenborn. Measurement of residual dipolar couplings of macromolecules aligned in the nematic phase of a colloidal suspension of rod-shaped viruses. *J Am Chem Soc*, 120(40):10571–10572, 1998. (cited p. 34.)

[75] M. R. Hansen, L. Mueller, and A. Pardi. Tunable alignment of macromolecules by filamentous phage yields dipolar coupling interactions. *Nat Struct Biol*, 5 (12):1065–74, 1998. (cited p. 34 and 147.)

[76] M. R. Hansen, P. Hanson, and A. Pardi. Filamentous bacteriophage for aligning rna, dna, and proteins for measurement of nuclear magnetic resonance dipolar coupling interactions. *Meth Enzymol*, 317:220, 2000. (cited p. 34 and 147.)

[77] J. Torbet and G. Maret. Fibres of highly oriented pf1 bacteriophage produced in a strong magnetic field. *J. Mol. Biol.*, 134:843–845, 1979. (cited p. 34.)

[78] M. Zweckstetter and A. Bax. Characterization of molecular alignment in aqueous suspensions of pf1 bacteriophage. *J Biomol NMR*, 20(4):365–77, 2001. (cited p. 34.)

[79] M. Rückert and G. Otting. Alignment of biological macromolecules in novel nonionic liquid crystalline media for nmr experiments. *J Am Chem Soc*, 122 (32):7793–7797, 2000. (cited p. 34 and 147.)

[80] R. Prosser, J. A. Losonczi, and I. V. Shiyanovskaya. Use of a novel aqueous liquid crystalline medium for high-resolution nmr of macromolecules in solution. *J Am Chem Soc*, 120(42):11010–11011, 1998.

[81] L. G. Barrientos, C. Dolan, and A. M. Gronenborn. Characterization of surfactant liquid crystal phases suitable for molecular alignment and measurement of dipolar couplings. *J Biomol NMR*, 16(4):329–37, 2000. (cited p. 34.)

[82] B. Koenig, J. Hu, M. Ottiger, S. Bose, R. Hendler, and A. Bax. Nmr measurement of dipolar couplings in proteins aligned by transient binding to purple membrane fragments. *J Am Chem Soc*, 121(6):1385–1386, 1999. (cited p. 34.)

[83] B. W. Koenig, D. C. Mitchell, S. König, S. Grzesiek, B. J. Litman, and A. Bax. Measurement of dipolar couplings in a transducin peptide fragment weakly bound to oriented photo-activated rhodopsin. *J Biomol NMR*, 16(2):121–5, 2000. (cited p. 34.)

[84] X. Dong, T. Kimura, J. Revol, and D. Gray. Effects of ionic strength on the isotropicâ chiral nematic phase transition of suspensions of cellulose crystallites. *Langmuir*, 12(8):2076–2082, 1996. (cited p. 34.)

[85] K. Fleming, D. Gray, S. Prasannan, and S. Matthews. Cellulose crystallites: A new and robust liquid crystalline medium for the measurement of residual dipolar couplings. *J Am Chem Soc*, 122(21):5224–5225, 2000.

[86] A. Denisov, E. Kloser, D. Gray, and A. K. Mittermaier. Protein alignment using cellulose nanocrystals: practical considerations and range of application. *J Biomol NMR*, 47:195–204, 2010. (cited p. 34.)

[87] J. L. Lorieau, L. Yao, and A. Bax. Liquid crystalline phase of g-tetrad dna for nmr study of detergent-solubilized proteins. *J Am Chem Soc*, 130(24):7536–7, 2008. (cited p. 34.)

[88] J. Ma, G. Goldberg, and N. Tjandra. Weak alignment of biomacromolecules in collagen gels: An alternative way to yield residual dipolar couplings for nmr measurements. *J Am Chem Soc*, 130(48):16148–16149, 2008. (cited p. 34.)

[89] R. Tycko, F. Blanco, and Y. Ishii. Alignment of biopolymers in strained gels: A new way to create detectable dipoleâ dipole couplings in high-resolution biomolecular nmr. *J Am Chem Soc*, 122(38):9340–9341, 2000. (cited p. 34 and 148.)

[90] H.-J. Sass, G. Musco, S. Stahl, P. T. Wingfield, and S. Grzesiek. Solution nmr of proteins within polyacrylamide gels: diffusional properties and residual alignment by mechanical stress or embedding of oriented purple membranes. *J Biomol NMR*, 18(4):303–309, 2000. (cited p. 34, 35 and 148.)

[91] T. Cierpicki and J. H. Bushweller. Charged gels as orienting media for measurement of residual dipolar couplings in soluble and integral membrane proteins. *J Am Chem Soc*, 126(49):16259–16266, 2004. (cited p. 35.)

[92] S. Meier, D. Häussinger, and S. Grzesiek. Charged acrylamide copolymer gels as media for weak alignment. *J Biomol NMR*, 24(4):351–356, 2002. (cited p. 35.)

[93] K. Ruan and J. R. Tolman. Composite alignment media for the measurement of independent sets of nmr residual dipolar couplings. *J Am Chem Soc*, 127 (43):15032–15033, 2005. (cited p. 35, 80 and 81.)

[94] E. Gebel, K. Ruan, J. R. Tolman, and D. Shortle. Multiple alignment tensors from a denatured protein. *J Am Chem Soc*, 128(29):9310–9311, 2006. (cited p. 35.)

[95] C. Cohen-Tannoudji, B. Diu, and F. Laloe. Mécanique quantique ii. *Hermann*, 1973. (cited p. 35, 36 and 37.)

[96] E. Henry and A. Szabo. Influence of vibrational motion on solid state line shapes and nmr relaxation. *J. Chem. Phys.*, 82:4753–4761, 1985. (cited p. 38.)

[97] L. Yao, B. Vögeli, J. Ying, and A. Bax. Nmr determination of amide n-h equilibrium bond length from concerted dipolar coupling measurements. *J Am Chem Soc*, 130(49):16518–20, 2008. (cited p. 38 and 82.)

[98] A. Saupe. Recent results in the field of liquid crystals. *Angew. Chem. Int. Ed.*, 7:97–112, 1968. (cited p. 40.)

[99] M. E. Rose. Elementary theory of angular momentum. *John Wiley & Sons*, 1957. (cited p. 42.)

[100] J.-C. Hus, L. Salmon, G. Bouvignies, J. Lotze, M. Blackledge, and R. Brüschweiler. 16-fold degeneracy of peptide plane orientations from residual dipolar couplings: Analytical treatment and implications for protein structure determination. *J Am Chem Soc*, 130(47):15927–15937, 2008. (cited p. 47 and 49.)

[101] H. M. Al-Hashimi, H. Valafar, M. Terrell, E. R. Zartler, M. K. Eidsness, and J. H. Prestegard. Variation of molecular alignment as a means of resolving orientational ambiguities in protein structures from dipolar couplings. *J Magn Reson*, 143(2):402–6, 2000. (cited p. 47.)

[102] G. Bouvignies, S. Meier, S. Grzesiek, and M. Blackledge. Ultrahigh-resolution backbone structure of perdeuterated protein gb1 using residual dipolar couplings from two alignment media. *Angew. Chem. Int. Ed.*, 45(48):8166–8169, 2006. (cited p. 50.)

[103] P. Bernado, L. Blanchard, P. Timmins, D. Marion, R. W. H. Ruigrok, and M. Blackledge. A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering. *Proc Natl Acad Sci USA*, 102(47):17002–7, 2005. (cited p. 51 and 197.)

[104] J. Meiler, J. J. Prompers, W. Peti, C. Griesinger, and R. Brüschweiler. Model-free approach to the dynamic interpretation of residual dipolar couplings in globular proteins. *J Am Chem Soc*, 123(25):6098–6107, 2001. (cited p. 52, 70, 71 and 88.)

[105] R. Brüschweiler and P. E. Wright. Nmr order parameters of biomolecules: A new analytical representation and application to the gaussian axial fluctuation model. *J Am Chem Soc*, 116(18):8426–8427, 1994. (cited p. 54.)

[106] R. Zwanzig. Nonequilibrium statistical mechanics. *Oxford University Press*, 2001. (cited p. 54.)

[107] D. A. McQuarrie. Statistical mechanics. *University Science Books*, 2000. (cited p. 54, 62 and 106.)

[108] I. S. Gradshtein, I. M. Ryzhik, A. Jeffrey, and D. Zwillinger. Table of integrals, series and products. *Academic Press*, 2007. (cited p. 55.)

[109] L. Frydman, T. Scherf, and A. Lupulescu. The acquisition of multidimensional nmr spectra within a single scan. *Proc Natl Acad Sci USA*, 99(25):15858, 2002. (cited p. 61.)

[110] P. Schanda and B. Brutscher. Very fast two-dimensional nmr spectroscopy for real-time investigation of dynamic events in proteins on the time scale of seconds. *J Am Chem Soc*, 127(22):8014–8015, 2005. (cited p. 61 and 146.)

[111] J. R. Tolman, J. M. Flanagan, M. A. Kennedy, and J. H. Prestegard. Nmr evidence for slow collective motions in cyanometmyoglobin. *Nat Struct Biol*, 4(4):292–7, 1997. (cited p. 62 and 65.)

[112] T. S. Ulmer, B. E. Ramirez, F. Delaglio, and A. Bax. Evaluation of backbone proton positions and dynamics in a small protein by liquid crystal nmr spectroscopy. *J Am Chem Soc*, 125(30):9179–91, 2003. (cited p. 62 and 127.)

[113] D. S. Berkholz, M. Shapovalov, R. D. Jr, and P. Karplus. Conformation dependence of backbone geometry in proteins. *Structure*, 17(10):1316–1325, 2009. (cited p. 63.)

[114] J.-C. Hus and R. Brüschweiler. Principal component method for assessing structural heterogeneity across multiple alignment media. *J Biomol NMR*, 24 (2):123–32, 2002. (cited p. 63 and 64.)

[115] J.-C. Hus, W. Peti, C. Griesinger, and R. Brüschweiler. Self-consistency analysis of dipolar couplings in multiple alignments of ubiquitin. *J Am Chem Soc*, 125(19):5596–5597, 2003. (cited p. 63, 64 and 71.)

[116] J. R. Tolman and K. Ruan. Nmr residual dipolar couplings as probes of biomolecular dynamics. *Chem Rev*, 106(5):1720–36, 2006. (cited p. 64.)

[117] M. Fischer, J. A. Losonczi, J. Weaver, and J. Prestegard. Domain orientation and dynamics in multidomain proteins from residual dipolar couplings. *Biochemistry*, 38(28):9013–9022, 1999. (cited p. 65.)

[118] D. Braddock, M. Cai, J. Baber, Y. Huang, and G. Clore. Rapid identification of medium-to large-scale interdomain motion in modular proteins using dipolar couplings. *J Am Chem Soc*, 123(35):8634–8635, 2001. (cited p. 65.)

[119] T. Ulmer, J. Werner, and I. Campbell. Sh3-sh2 domain orientation in src kinases: Nmr studies of fyn. *Structure*, 10(7):901–911, 2002.

[120] I. Bertini, C. D. Bianco, I. Gelis, N. Katsaros, C. Luchinat, G. Parigi, M. Peana, A. Provenzani, and M. Zoroddu. Experimentally exploring the conformational space sampled by domain reorientation in calmodulin. *Proc Natl Acad Sci USA*, 101(18):6841, 2004. (cited p. 65.)

[121] H. M. Al-Hashimi, Y. Gosser, A. Gorin, W. Hu, A. Majumdar, and D. Patel. Concerted motions in hiv-1 tar rna may allow access to bound state conformations: Rna dynamics from nmr residual dipolar couplings. *J. Mol. Biol.*, 315(2):95–102, 2002. (cited p. 65.)

[122] Q. Zhang, R. Throolin, S. W. Pitt, A. Serganov, and H. M. Al-Hashimi. Probing motions between equivalent rna domains using magnetic field induced residual dipolar couplings: accounting for correlations between motions and alignment. *J Am Chem Soc*, 125(35):10530–10531, 2003. (cited p. 65.)

[123] F. Tian, H. M. Al-Hashimi, J. Craighead, and J. H. Prestegard. Conformational analysis of a flexible oligosaccharide using residual dipolar couplings. *J Am Chem Soc*, 123(3):485–492, 2001. (cited p. 65 and 66.)

[124] N. Skrynnikov. Orienting molecular fragments and molecules with residual dipolar couplings. *C. R. Physique*, 5(3):359–375, 2004. (cited p. 65.)

[125] N. Skrynnikov, N. Goto, D. Yang, W. Choy, J. R. Tolman, G. Mueller, and L. Kay. Orienting domains in proteins using dipolar couplings measured by liquid-state nmr: differences in solution and crystal forms of maltodextrin binding protein loaded with beta-cyclodextrin. *J. Mol. Biol.*, 295(5):1265–1273, 2000. (cited p. 65.)

[126] E. de Alba, J. Baber, and N. Tjandra. The use of residual dipolar coupling in concert with backbone relaxation rates to identify conformational exchange by nmr. *J Am Chem Soc*, 121(17):4282–4283, 1999. (cited p. 65.)

[127] J. Chou, S. Li, C. Klee, and A. Bax. Solution structure of ca2+–calmodulin reveals flexible hand-like properties of its domains. *Nat Struct Mol Biol*, 8(11): 990–997, 2001.

[128] F. Mareuil, C. Sizun, J. Perez, M. Schoenauer, J.-Y. Lallemand, and F. Bontems. A simple genetic algorithm for the optimization of multidomain protein homology models driven by nmr residual dipolar coupling and small angle x-ray scattering data. *Eur Biophys J*, 37(1):95–104, 2007. (cited p. 65.)

[129] J. R. Tolman, H. M. Al-Hashimi, L. E. Kay, and J. H. Prestegard. Structural and dynamic analysis of residual dipolar coupling data for proteins. *J Am Chem Soc*, 123(7):1416–24, 2001. (cited p. 65.)

[130] R. Kaptein, E. Zuiderweg, and R. Scheek. A protein structure from nuclear magnetic resonance data : lac repressor headpiece. *J. Mol. Biol.*, 182(1):179, 1985. (cited p. 66.)

[131] G. M. Clore, A. M. Gronenborn, A. T. Brünger, and M. Karplus. Solution conformation of a heptadecapeptide comprising the dna binding helix f of the cyclic amp receptor protein of escherichia coli. combined use of 1h nuclear magnetic resonance and restrained molecular dynamics. *J. Mol. Biol.*, 186(2): 435–55, 1985.

[132] E. R. Zuiderweg, R. M. Scheek, R. Boelens, W. F. van Gunsteren, and R. Kaptein. Determination of protein structures from nuclear magnetic resonance data using a restrained molecular dynamics approach: the lac repressor dna binding domain. *Biochimie*, 67(7-8):707–15, 1985.

[133] Clore, A. Brünger, M. Karplus, and A. Gronenborn. Application of molecular dynamics with interproton distance restraints to three-dimensional protein structure determination : A model study of crambin. *J. Mol. Biol.*, 191(4):523, 1986.

[134] G. M. Clore, M. Nilges, D. K. Sukumaran, A. T. Brünger, M. Karplus, and A. M. Gronenborn. The three-dimensional structure of alpha1-purothionin in solution: combined use of nuclear magnetic resonance, distance geometry and restrained molecular dynamics. *EMBO J*, 5(10):2729–35, 1986. (cited p. 66.)

[135] G. M. Clore and C. Schwieters. Amplitudes of protein backbone dynamics and correlated motions in a small alpha/beta protein: Correspondence of dipolar coupling and heteronuclear relaxation measurements. *Biochemistry*, 43(33):10678–10691, 2004. (cited p. 66 and 67.)

[136] G. M. Clore and C. Schwieters. How much backbone motion in ubiquitin is required to account for dipolar coupling data measured in multiple alignment media as assessed by independent cross-validation? *J Am Chem Soc*, 126(9): 2923–2938, 2004. (cited p. 67.)

[137] C. Schwieters and G. M. Clore. A physical picture of atomic motions within the dickerson dna dodecamer in solution derived from joint ensemble refinement against nmr and large-angle x-ray scattering dataâ . *Biochemistry*, 46(5): 1152–1166, 2007. (cited p. 67.)

[138] K. Lindorff-Larsen, R. B. Best, M. A. Depristo, C. M. Dobson, and M. Vendruscolo. Simultaneous determination of protein structure and dynamics. *Nature*, 433(7022):128–132, 2005. (cited p. 67.)

[139] O. F. Lange, N.-A. Lakomek, C. Fares, G. F. Schroder, K. F. A. Walter, S. Becker, J. Meiler, H. Grubmüller, C. Griesinger, and B. L. D. Groot. Recognition dynamics up to microseconds revealed from an rdc-derived ubiquitin ensemble in solution. *Science*, 320(5882):1471–5, 2008. (cited p. 67 and 114.)

[140] J. Huang and S. Grzesiek. Ensemble calculations of unstructured proteins constrained by rdc and pre data: A case study of urea-denatured ubiquitin. *J Am Chem Soc*, 132(2):694–705, 2010. (cited p. 67, 226 and 233.)

[141] Q. Zhang, A. Stelzer, C. Fisher, and H. M. Al-Hashimi. Visualizing spatially correlated dynamics that directs rna conformational transitions. *Nature*, 450 (7173):1263–1267, 2007. (cited p. 67.)

[142] S. A. Adcock and J. A. McCammon. Molecular dynamics: survey of methods for simulating the activity of proteins. *Chem Rev*, 106(5):1589–1615, 2006. (cited p. 67, 68, 108 and 152.)

[143] M. Karplus and J. McCammon. Molecular dynamics simulations of biomolecules. *Nat Struct Mol Biol*, 9(9):646–652, 2002. (cited p. 67.)

[144] J. Ponder and D. Case. Force fields for protein simulations. *Advances in protein chemistry*, 66:27–85, 2003. (cited p. 67.)

[145] A. Voter. Hyperdynamics: Accelerated molecular dynamics of infrequent events. *Phys Rev Lett*, 78(20):3908–3911, 1997. (cited p. 68.)

[146] B. Berne and J. Straub. Novel methods of sampling phase space in the simulation of biological systems. *Curr. Opin. Struct. Biol.*, 7(2):181–189, 1997.

[147] R. Elber. Long-timescale simulation methods. *Curr. Opin. Struct. Biol.*, 15(2): 151–156, 2005. (cited p. 68.)

[148] S. A. Showalter and R. Brüschweiler. Quantitative molecular ensemble interpretation of nmr dipolar couplings without restraints. *J Am Chem Soc*, 129 (14):4158–9, 2007. (cited p. 68, 97 and 259.)

[149] S. F. Lienin, T. Bremi, B. Brutscher, R. Brüschweiler, and R. R. Ernst. Anisotropic intramolecular backbone dynamics of ubiquitin characterized by nmr relaxation and md computer simulation. *J Am Chem Soc*, 120(38): 9870–9879, 1998. (cited p. 68, 93, 96, 99 and 267.)

[150] Y. Chen, S. Campbell, and N. Dokholyan. Deciphering protein dynamics from nmr data using explicit structure sampling and selection. *Biophys J*, 93 (7):2300–2306, 2007. (cited p. 68 and 69.)

[151] J. A. Marsh and J. D. Forman-Kay. Structure and disorder in an unfolded state under nondenaturing conditions from ensemble models consistent with a large number of experimental restraints. *J. Mol. Biol.*, 391(2):359–374, 2009. (cited p. 69.)

[152] A. C. Stelzer, A. T. Frank, M. H. Bailor, I. Andricioaei, and H. M. Al-Hashimi. Constructing atomic-resolution rna structural ensembles using md and motionally decoupled nmr rdcs. *Methods*, 49(2):167–173, 2009. (cited p. 69.)

[153] N.-A. Lakomek, K. F. A. Walter, C. Fares, O. F. Lange, B. L. Groot, H. Grubmüller, R. Brüschweiler, A. Munk, S. Becker, J. Meiler, and C. Griesinger. Self-consistent residual dipolar coupling based model-free analysis for the robust determination of nanosecond to microsecond protein dynamics. *J Biomol NMR*, 41(3):139–155, 2008. (cited p. 70, 71, 80, 81, 98 and 99.)

[154] W. Peti, J. Meiler, R. Brüschweiler, and C. Griesinger. Model-free analysis of protein backbone motion from residual dipolar couplings. *J Am Chem Soc*, 124(20):5822–33, 2002. (cited p. 71.)

[155] N.-A. Lakomek, C. Fares, S. Becker, T. Carlomagno, J. Meiler, and C. Griesinger. Side-chain orientation and hydrogen-bonding imprint supra-tau(c) motion on the protein backbone of ubiquitin. *Angew. Chem. Int. Ed.*, 44 (47):7776–8, 2005.

[156] N. Lakomek, T. Carlomagno, S. Becker, C. Griesinger, and J. Meiler. A thorough dynamic interpretation of residual dipolar couplings in ubiquitin. *J Biomol NMR*, 34(2):101–115, 2006.

[157] N.-A. Lakomek, T. Carlomagno, S. Becker, C. Griesinger, and J. Meiler. A thorough dynamic interpretation of residual dipolar couplings in ubiquitin. *J Biomol NMR*, 34(2):101–115, 2006. (cited p. 71, 80 and 81.)

[158] J. R. Tolman. A novel approach to the retrieval of structural and dynamic information from residual dipolar couplings using several oriented media in biomolecular nmr spectroscopy. *J Am Chem Soc*, 124(40):12020–30, 2002. (cited p. 71 and 72.)

[159] K. B. Briggman and J. R. Tolman. De novo determination of bond orientations and order parameters from residual dipolar couplings with high accuracy. *J Am Chem Soc*, 125(34):10164–10165, 2003. (cited p. 72, 80 and 81.)

[160] L. Yao, B. Vogeli, D. Torchia, and A. Bax. Simultaneous nmr study of protein structure and dynamics using conservative mutagenesis. *J. Phys. Chem. B*, 112 (19):6045–6056, 2008. (cited p. 72.)

[161] P. Bernado and M. Blackledge. Anisotropic small amplitude peptide plane dynamics in proteins from residual dipolar couplings. *J Am Chem Soc*, 126 (15):4907–4920, 2004. (cited p. 74 and 135.)

[162] P. Bernado and M. Blackledge. Local dynamic amplitudes on the protein backbone from dipolar couplings: toward the elucidation of slower motions in biomolecules. *J Am Chem Soc*, 126(25):7760–7761, 2004. (cited p. 75.)

[163] G. Bouvignies, P. Bernado, and M. Blackledge. Protein backbone dynamics from n-hn dipolar couplings in partially aligned systems: a comparison of motional models in the presence of structural noise. *J Magn Reson*, 173(2): 328–38, 2005. (cited p. 75 and 93.)

[164] G. Bouvignies, P. Bernado, S. Meier, K. Cho, S. Grzesiek, R. Brüschweiler, and M. Blackledge. Identification of slow correlated motions in proteins using residual dipolar and hydrogen-bond scalar couplings. *Proc Natl Acad Sci USA*, 102(39):13885–90, 2005. (cited p. 75, 93, 122, 142 and 261.)

[165] J. Hus, D. Marion, and M. Blackledge. Determination of protein backbone structure using only residual dipolar couplings. *J Am Chem Soc*, 123(7): 1541–1542, 2001. (cited p. 76 and 197.)

[166] G. Bouvignies, P. R. L. Markwick, and M. Blackledge. Characterization of protein dynamics from residual dipolar couplings using the three dimensional gaussian axial fluctuation model. *Proteins*, 71(1):353–63, 2008. (cited p. 76, 80, 88 and 93.)

[167] G. Bouvignies, P. R. L. Markwick, R. Brüschweiler, and M. Blackledge. Simultaneous determination of protein backbone structure and dynamics from residual dipolar couplings. *J Am Chem Soc*, 128(47):15100–15101, 2006. (cited p. 82, 93, 122, 127, 130, 131, 134, 137 and 142.)

[168] P. Holland and R. Welsch. Robust regression using iteratively reweighted least-squares. *Commun Stat Theory*, A6:813–827, 1977. (cited p. 83.)

[169] R. Eberhart and J. Kennedy. A new optimizer using particle swarm theory. *Proceedings of the Sixth International Symposium on Micromachine and Human Science*, 1995. (cited p. 83.)

[170] L. Kang, Z. Cai, X. Yan, and Y. Liu. Advances in computation and intelligenceâ: Third international symposium on intelligence computation and applications. *Springer*, 2008. (cited p. 83.)

[171] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. Numerical recipes in c: the art of scientific computing. *Cambridge University Press*, 1988. (cited p. 84 and 257.)

[172] P. Rousseeuw and C. Croux. Alternatives to the median absolute deviation. *J Am Stat Assoc*, 88(424):1273–1283, 1993. (cited p. 85.)

[173] H. Motulsky and A. Christopoulos. Fitting models to biological data using linear and nonlinear regression. *Oxford University Press*, page 351, 2004. (cited p. 86 and 87.)

[174] B. Clarke, E. Fokoué, and H. H. Zhang. Principles and theory for data mining and machine learning. *Springer*, page 781, 2009. (cited p. 87.)

[175] G. Cornilescu, J. Marquardt, M. Ottiger, and A. Bax. Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase. *J Am Chem Soc*, 120(27):6836–6837, 1998. (cited p. 93, 101, 114 and 177.)

[176] S. Chang and N. Tjandra. Temperature dependence of protein backbone motion from carbonyl 13c and amide 15n nmr relaxation. *J Magn Res*, 174(1): 43–53, 2005. (cited p. 99 and 118.)

[177] F. Cordier and S. Grzesiek. Temperature-dependence of protein hydrogen bond properties as studied by high-resolution nmr. *J. Mol. Biol.*, 317(5):739–52, 2002. (cited p. 103 and 118.)

[178] D. Hamelberg, J. Mongan, and J. A. McCammon. Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules. *J. Chem. Phys.*, 120:11919–11929, 2004. (cited p. 105 and 106.)

[179] D. Hamelberg, T. Shen, and J. A. McCammon. Phosphorylation effects on cis/trans isomerization and the backbone conformation of serineâ proline motifs: Accelerated molecular dynamics analysis. *J Am Chem Soc*, 127(6): 1969–1974, 2005. (cited p. 105.)

[180] D. Hamelberg and J. A. McCammon. Fast peptidyl cisâ trans isomerization within the flexible gly-rich flaps of hiv-1 protease. *J Am Chem Soc*, 127(40): 13778–13779, 2005.

[181] P. R. L. Markwick, G. Bouvignies, and M. Blackledge. Exploring multiple timescale motions in protein gb3 using accelerated molecular dynamics and nmr spectroscopy. *J Am Chem Soc*, 129(15):4724–30, 2007. (cited p. 105, 106, 108, 114 and 131.)

[182] B. Diu, D. Lederer, and B. Roulet. Éléments de physique statistique. *Hermann*, 1989. (cited p. 106.)

[183] H. Berendsen, J. Postma, W. vangunsteren, A. Dinola, and J. Haak. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81(8):3684–3690. (cited p. 108 and 152.)

[184] Cheatham, J. Miller, T. Fox, T. Darden, and P. A. Kollman. Molecular dynamics simulations on solvated biomolecular systems: the particle mesh ewald method leads to stable trajectories of dna, rna, and proteins. *J Am Chem Soc*, 117(4193):4194, 1995. (cited p. 108 and 152.)

[185] D. Case, T. Darden, T. C. III, C. Simmerling, J. Wang, R. Duke, R. Luo, K. Merz, B. Wang, D. Pearlman, M. Crowley, S. Brozell, V. Tsui, H. Gohlke, J. Mongan, V. Hornak, G. Cui, P. Beroza, C. Schafmeister, J. Cadwell, W. Ross, and P. A. Kollman. Amber 8. *University of California, San Francisco*, 2004. (cited p. 108.)

[186] S. Vijay-kumar, C. Bugg, and W. Cook. Structure of ubiquitin refined at 1.8 å resolution. *J. Mol. Biol.*, 194:531–544, 1987. (cited p. 114.)

[187] B. Brooks and M. Karplus. Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proc Natl Acad Sci USA*, 80(21):6571, 1983. (cited p. 122.)

[188] C. Pickover. Spectrographic representation of globular protein breathing motions. *Science*, 223(4632):181, 1984.

[189] D. Case. Normal mode analysis of protein dynamics. *Curr. Opin. Struct. Biol.*, 4(2):285–290, 1994. (cited p. 122 and 133.)

[190] A. Dhulesia, D. Abergel, and G. Bodenhausen. Networks of coupled rotators: Relationship between structures and internal dynamics in metal-binding proteins. applications to apo-and holo-calbindin. *J Am Chem Soc*, 129(16): 4998–5006, 2007. (cited p. 122.)

[191] J. R. Lewandowski, J. Sein, H.-J. Sass, S. Grzesiek, M. Blackledge, and L. Emsley. Measurement of site-specific 13c spinâ lattice relaxation in a crystalline protein. *J Am Chem Soc*, pages 3055–3079, 2010. (cited p. 122.)

[192] W. P. Kelly and M. P. H. Stumpf. Protein–proteininteractions: fromglobal tolocal analyses. *Current Opinion in Biotechnology*, 19:396–403, 2008. (cited p. 146.)

[193] R. B. Russell, F. Alber, P. Aloy, F. P. Davis, D. Korkin, M. Pichaud, M. Topf, and A. Sali. A structural perspective on protein–protein interactions. *Curr. Opin. Struct. Biol.*, 14:313–324, 2004. (cited p. 146.)

[194] F. Delaglio, S. Grzesiek, G. W. Vuister, G. Zhu, J. Pfeifer, and A. Bax. Nmrpipe: a multidimensional spectral processing system based on unix pipes. *J Biomol NMR*, 6(3):277–93, 1995. (cited p. 146.)

[195] T. Goddard and D. Kneller. *University of California*, 2003. (cited p. 146.)

[196] J. L. Ortega-Roldan, M. R. Jensen, B. Brutscher, A. I. Azuaga, M. Blackledge, and N. A. J. van Nuland. Accurate characterization of weak macromolecular interactions by titration of nmr residual dipolar couplings: application to the cd2ap sh3-c:ubiquitin complex. *Nucleic Acids Res*, 37(9):e70, 2009. (cited p. 146, 170, 172, 181 and 185.)

[197] N. A. Farrow, R. Muhandiram, A. U. Singer, S. M. Pascal, C. M. Kay, G. Gish, S. E. Shoelson, T. Pawson, J. D. Forman-Kay, and L. E. Kay. Backbone dynamics of a free and a phosphopeptide-complexed src homology 2 domain studied by 15n nmr relaxation. *Biochemistry*, 33(19):5984–6003, 1994. (cited p. 146.)

[198] P. Schanda, H. V. Melckebeke, and B. Brutscher. Speeding up three-dimensional protein nmr experiments to a few minutes. *J Am Chem Soc*, 128(28):9042–9043, 2006. (cited p. 146.)

[199] E. Lescop, P. Schanda, and B. Brutscher. A set of best triple-resonance experiments for time-optimized protein resonance assignment. *J Magn Res*, 187(1):163–169, 2007. (cited p. 146.)

[200] J. Hus, D. Marion, and M. Blackledge. De novo determination of protein structure by nmr using orientational and long-range order restraints. *J. Mol. Biol.*, 298(5):927–936, 2000. (cited p. 151.)

[201] A. Brunger, P. Adams, G. Clore, W. DeLano, P. Gros, R. Grosse-Kunstleve, J. Jiang, J. Kuszewski, M. Nilges, N. Pannu, R. J. Read, L. M. Rice, T. Simonson, and G. L. Warren. Crystallography & nmr system: a new software suite for macromolecular structure determination. *Acta Crystallographica Section D: Biological Crystallography*, 54(5):905–921, 1998. (cited p. 151.)

[202] A. Brunger. Version 1.2 of the crystallography and nmr system. *Nature protocols*, 2(11):2728–2733, 2007. (cited p. 151.)

[203] J. L. Ortega-Roldan, M. L. R. Romero, A. Ora, E. Ab, O. L. Mayorga, A. I. Azuaga, and N. A. J. van Nuland. The high resolution nmr structure of the third sh3 domain of cd2ap. *J Biomol NMR*, 39(4):331–6, 2007. (cited p. 151, 166 and 261.)

[204] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl. Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput*, 4(3):435–447, 2008. (cited p. 152.)

[205] X. Wu, B. Knudsen, S. Feller, J. Zheng, A. Sali, D. Cowburn, H. Hanafusa, and J. Kuriyan. Structural basis for the specific interaction of lysine-containing proline-rich peptides with the n-terminal sh3 domain of c-crk. *Structure*, 3(2): 215–226, 1995. (cited p. 154.)

[206] V. Chevelkov, K. Faelber, A. Diehl, U. Heinemann, H. Oschkinat, and B. Reif. Detection of dynamic water molecules in a microcrystalline sample of the sh3 domain of alpha-spectrin by mas solid-state nmr. *J Biomol NMR*, 31(4): 295–310, 2005. (cited p. 154.)

[207] D. Jozic, N. Cárdenes, Y. Deribe, G. Moncalián, D. Hoeller, Y. Groemping, I. Dikic, K. Rittinger, and J. Bravo. Cbl promotes clustering of endocytic adaptor proteins. *Nat Struct Mol Biol*, 12(11):972–979, 2005. (cited p. 154.)

[208] G. Moncalián, N. Cárdenes, Y. L. Deribe, M. Spínola-Amilibia, I. Dikic, and J. Bravo. Atypical polyproline recognition by the cms n-terminal src homology 3 domain. *J Biol Chem*, 281(50):38845–53, 2006. (cited p. 154.)

[209] J. Vaynberg and J. Qin. Weak protein–protein interactions as probed by nmr spectroscopy. *TRENDS in Biotechnology*, 24(1):22–27, 2006. (cited p. 165.)

[210] J. Iwahara and G. M. Clore. Detecting transient intermediates in macromolecular binding by paramagnetic nmr. *Nature*, 440:1227–1230, 2006. (cited p. 165.)

[211] E. Olejniczak, R. Meadows, H. Wang, M. Cai, D. Nettesheim, and S. Fesik. Improved nmr structures of protein/ligand complexes using residual dipolar couplings. *J Am Chem Soc*, 121:9249–9250, 1999. (cited p. 165.)

[212] D. S. Garrett, Y. J. Seok, A. Peterkofsky, A. Gronenborn, and G. Clore. Solution structure of the 40,000 mr phosphoryl transfer complex between the n-terminal domain of enzyme i and hpr. *Nat Struct Biol*, 6:166–173, 1999.

[213] G. M. Clore. Accurate and rapid docking of protein–protein complexes on the basis of intermolecular nuclear overhauser enhancement data and dipolar couplings by rigid body minimization. *Proc Natl Acad Sci USA*, 97 (16):9021–9025, 2000. (cited p. 165.)

[214] K. Sugase, J. Lansing, H. Dyson, and P. E. Wright. Tailoring relaxation dispersion experiments for fast-associating protein complexes. *J Am Chem Soc*, 129(44):13406–13407, 2007. (cited p. 168.)

[215] J. Carver and R. Richards. A general two-site solution for the chemical exchange produced dependence of t2 upon the carr-purcell pulse separation. *J Magn Res*, 6:89–105, 1972. (cited p. 168.)

[216] D. M. Korzhnev, I. Bezsonova, S. Lee, T. V. Chalikian, and L. E. Kay. Alternate binding modes for a ubiquitin-sh3 domain interaction studied by nmr spectroscopy. *J. Mol. Biol.*, 386(2):391–405, 2009. (cited p. 179.)

[217] A. Fink. Natively unfolded proteins. *Curr. Opin. Struct. Biol.*, 15(1):35–41, 2005. (cited p. 191 and 195.)

[218] A. Dunker, I. Silman, V. Uversky, and J. Sussman. Function and structure of inherently disordered proteins. *Curr. Opin. Struct. Biol.*, 18(6):756–764, 2008.

[219] P. Tompa. Intrinsically unstructured proteins. *Trends Biochem Sci*, 27(10): 527–533, 2002. (cited p. 191 and 192.)

[220] P. E. Wright and H. J. Dyson. Linking folding and binding. *Curr. Opin. Struct. Biol.*, 19(1):31–8, 2009. (cited p. 191.)

[221] D. Eliezer and A. G. Palmer. Biophysics: proteins hunt and gather. *Nature*, 447(7147):920–1, 2007. (cited p. 191.)

[222] P. Tompa and M. Fuxreiter. Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. *Trends Biochem Sci*, 33(1):2–8, 2008. (cited p. 191.)

[223] D. Eliezer. Biophysical characterization of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.*, 19(1):23–30, 2009. (cited p. 192.)

[224] T. Keiderling and Q. Xu. Unfolded peptides and proteins studied with infrared absorption and vibrational circular dichroism spectra. *Advances in protein chemistry*, 62:111–161, 2002. (cited p. 192.)

[225] V. Uversky. Natively unfolded proteins: a point where biology waits for physics. *Protein Sci*, 11(4):739–756, 2002. (cited p. 192.)

[226] I. Millett, S. Doniach, and K. Plaxco. Toward a taxonomy of the denatured state: small angle scattering studies of unfolded proteins. *Advances in protein chemistry*, 62:241–262, 2002. (cited p. 192.)

[227] P. Bernado, E. Mylonas, M. V. Petoukhov, M. Blackledge, and D. I. Svergun. Structural characterization of flexible proteins using small-angle x-ray scattering. *J Am Chem Soc*, 129(17):5656–64, 2007. (cited p. 192.)

[228] M. Fuxreiter, I. Simon, P. Friedrich, and P. Tompa. Preformed structural elements feature in partner recognition by intrinsically unstructured proteins. *J. Mol. Biol.*, 338(5):1015–1026, 2004. (cited p. 192.)

[229] M. Jensen, P. Markwick, S. Meier, C. Griesinger, M. Zweckstetter, S. Grzesiek, P. Bernado, and M. Blackledge. Quantitative determination of the conformational properties of partially folded and intrinsically disordered proteins using nmr dipolar couplings. *Structure*, 17(9):1169–1185, 2009. (cited p. 192.)

[230] H. Dyson and P. E. Wright. Unfolded proteins and protein folding studied by nmr. *Chem Rev*, 104(8):3607–3622, 2004. (cited p. 192, 193 and 223.)

[231] S. Spera and A. Bax. Empirical correlation between protein backbone conformation and c. alpha. and c. beta. 13c nuclear magnetic resonance chemical shifts. *J Am Chem Soc*, 113(14):5490–5492, 1991. (cited p. 193.)

[232] D. S. Wishart, B. D. Sykes, and F. Richards. The chemical shift index: a fast and simple method for the assignment of protein secondary structure through nmr spectroscopy. *Biochemistry*, 31(6):1647–1651, 1992. (cited p. 193.)

[233] D. S. Wishart, C. Bigam, A. Holm, R. Hodges, and B. D. Sykes. 1 h, 13 c and 15 n random coil nmr chemical shifts of the common amino acids. i. investigations of nearest-neighbor effects. *J Biomol NMR*, 5(1):67–81, 1995. (cited p. 193 and 216.)

[234] S. Schwarzinger, G. Kroon, T. Foss, J. Chung, P. E. Wright, and H. Dyson. Sequence-dependent correction of random coil nmr chemical shifts. *J Am Chem Soc*, 123(13):2970–2978, 2001.

[235] Y. Wang and O. Jardetzky. Probability-based protein secondary structure identification using combined nmr chemical-shift data. *Protein Sci*, 11(4): 852–861, 2002. (cited p. 193.)

[236] L. Wang, H. Eghbalnia, A. Bahrami, and J. Markley. Linear analysis of carbon-13 chemical shift differences and its application to the detection and correction of errors in referencing and spin system identifications. *J Biomol NMR*, 32(1):13–22, 2005. (cited p. 193.)

[237] J. Marsh, V. Singh, Z. Jia, and J. Forman-Kay. Sensitivity of secondary structure propensities to sequence differences between a-and g-synuclein: implications for fibrillation. *Protein Sci*, 15(12):2795–2804, 2006. (cited p. 193 and 216.)

[238] L. Serrano. Comparison between the [phi] distribution of the amino acids in the protein database and nmr data indicates that amino acids have various [phi] propensities in the random coil conformation. *J. Mol. Biol.*, 254(2): 322–333, 1995. (cited p. 193.)

[239] L. Smith, K. Bolin, H. Schwalbe, M. MacArthur, J. Thornton, and C. Dobson. Analysis of main chain torsion angles in proteins: prediction of nmr coupling constants for native and random coil conformations. *J. Mol. Biol.*, 255(3): 494–506, 1996. (cited p. 193.)

[240] A. Buevich and J. Baum. Dynamics of unfolded proteins: Incorporation of distributions of correlation times in the model free analysis of nmr relaxation data. *J Am Chem Soc*, 121(37):8671–8672, 1999. (cited p. 193.)

[241] M. Tollinger, N. Skrynnikov, F. Mulder, J. Forman-Kay, and L. Kay. Slow dynamics in folded and unfolded states of an sh3 domain. *J Am Chem Soc*, 123(46):11341–11352, 2001. (cited p. 193.)

[242] J. Klein-Seetharaman, M. Oikawa, S. Grimshaw, J. Wirmer, E. Duchardt, T. Ueda, T. Imoto, L. Smith, C. Dobson, and H. Schwalbe. Long-range interactions within a nonnative protein. *Science*, 295(5560):1719, 2002. (cited p. 193.)

[243] M. Zweckstetter and A. Bax. Prediction of sterically induced alignment in a dilute liquid crystalline phase: aid to protein structure determination by nmr. *J Am Chem Soc*, 122(15):3791–3792, 2000. (cited p. 194, 195, 197 and 203.)

[244] M. Louhivuori, K. Pääkkönen, K. Fredriksson, P. Permi, J. Lounila, and A. Annila. On the origin of residual dipolar couplings from denatured proteins. *J Am Chem Soc*, 125(50):15647–15650, 2003. (cited p. 194.)

[245] K. Fredriksson, M. Louhivuori, P. Permi, and A. Annila. On the interpretation of residual dipolar couplings as reporters of molecular dynamics. *J Am Chem Soc*, 126(39):12646–12650, 2004.

[246] O. Obolensky, K. Schlepckow, H. Schwalbe, and A. Solov'yov. Theoretical framework for nmr residual dipolar couplings in unfolded proteins. *J Biomol NMR*, 39(1):1–16, 2007. (cited p. 194.)

[247] R. Mohana-Borges, N. Goto, G. Kroon, H. Dyson, and P. E. Wright. Structural characterization of unfolded states of apomyoglobin using residual dipolar couplings. *J. Mol. Biol.*, 340(5):1131–1142, 2004. (cited p. 194.)

[248] W. Fieber, S. Kristjansdottir, and F. M. Poulsen. Short-range, long-range and transition state interactions in the denatured state of acbp from residual dipolar couplingsâ. *J. Mol. Biol.*, 339(5):1191–1199, 2004. (cited p. 194.)

[249] L. Skora, M.-K. Cho, H.-Y. Kim, S. Becker, C. O. Fernandez, M. Blackledge, and M. Zweckstetter. Charge-induced molecular alignment of intrinsically disordered proteins. *Angew. Chem. Int. Ed.*, 45(42):7012–5, 2006. (cited p. 195.)

[250] G. Clore, C. Tang, and J. Iwahara. Elucidating transient macromolecular interactions using paramagnetic relaxation enhancement. *Curr. Opin. Struct. Biol.*, 17(5):603–616, 2007. (cited p. 196 and 225.)

[251] A. Der-Sarkissian, C. Jao, J. Chen, and R. Langen. Structural organization of alpha-synuclein fibrils studied by site-directed spin labeling. *J Biol Chem*, 278 (39):37530, 2003.

[252] C. W. Bertoncini, Y. Jung, C. Fernandez, W. Hoyer, C. Griesinger, T. Jovin, and M. Zweckstetter. Release of long-range tertiary interactions potentiates aggregation of natively unstructured alpha-synuclein. *Proc Natl Acad Sci USA*, 102(5):1430, 2005. (cited p. 196, 224, 229 and 234.)

[253] J. Iwahara, C. Tang, and G. M. Clore. Practical aspects of 1h transverse paramagnetic relaxation enhancement measurements on macromolecules. *J Magn Res*, 184(2):185–195, 2007. (cited p. 196.)

[254] J. Gillespie and D. Shortle. Characterization of long-range structure in the denatured state of staphylococcal nuclease. i. paramagnetic relaxation enhancement by nitroxide spin labels. *J. Mol. Biol.*, 268(1):158–169, 1997. (cited p. 196.)

[255] Y. Xue, I. Podkorytov, D. Rao, N. Benjamin, H. Sun, and N. Skrynnikov. Paramagnetic relaxation enhancements in unfolded proteins: Theory and application to drkn sh3 domain. *Protein Sci*, 18(7):1401–1424, 2009. (cited p. 196.)

[256] M. Dedmon, K. Lindorff-Larsen, J. Christodoulou, M. Vendruscolo, and C. M. Dobson. Mapping long-range interactions in alpha-synuclein using spin-label nmr and ensemble molecular dynamics simulations. *J Am Chem Soc*, 127(2): 476–477, 2005. (cited p. 196 and 234.)

[257] J. Allison, P. Varnai, C. M. Dobson, and M. Vendruscolo. Determination of the free energy landscape of alpha-synuclein using spin label nuclear magnetic resonance measurements. *J Am Chem Soc*, 131(51):18314–18326, 2009. (cited p. 196 and 233.)

[258] D. Felitsky, M. Lietzow, H. Dyson, and P. E. Wright. Modeling transient collapsed states of an unfolded protein to provide insights into early folding events. *Proc Natl Acad Sci USA*, 105(17):6278, 2008. (cited p. 196.)

[259] M. Levitt. A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.*, 104:59, 1976. (cited p. 197.)

[260] A. Almond and J. B. Axelsen. Physical interpretation of residual dipolar couplings in neutral aligned media. *J Am Chem Soc*, 124(34):9986–9987, 2002. (cited p. 197.)

[261] M. Jensen, K. Houben, E. Lescop, L. Blanchard, R. Ruigrok, and M. Blackledge. Quantitative conformational analysis of partially folded proteins from residual dipolar couplings: application to the molecular recognition element of sendai virus nucleoprotein. *J Am Chem Soc*, 130(25):8055–8061, 2008. (cited p. 198, 216, 218, 220, 221 and 223.)

[262] M. Jensen and M. Blackledge. On the origin of nmr dipolar waves in transient helical elements of partially folded proteins. *J Am Chem Soc*, 130(34):11266–11267, 2008. (cited p. 198.)

[263] P. Bernado, C. W. Bertoncini, C. Griesinger, M. Zweckstetter, and M. Blackledge. Defining long-range order and local disorder in native alpha-synuclein using residual dipolar couplings. *J Am Chem Soc*, 127(51):17968–17969, 2005. (cited p. 198, 226 and 234.)

[264] M. Mukrasch, P. Markwick, J. Biernat, M. von Bergen, P. Bernado, C. Griesinger, E. Mandelkow, M. Zweckstetter, and M. Blackledge. Highly populated turn conformations in natively unfolded tau protein identified from residual dipolar couplings and molecular simulation. *J Am Chem Soc*, 129(16):5235–5243, 2007. (cited p. 198 and 201.)

[265] S. Meier, S. Grzesiek, and M. Blackledge. Mapping the conformational landscape of urea-denatured ubiquitin using residual dipolar couplings. *J Am Chem Soc*, 129(31):9799–9807, 2007. (cited p. 198, 201, 202, 203, 207, 211 and 221.)

[266] J. Marsh, J. M. R. Baker, M. Tollinger, and J. Forman-Kay. Calculation of residual dipolar couplings from disordered state ensembles using local alignment. *J Am Chem Soc*, 130(25):7804–7805, 2008. (cited p. 203 and 223.)

[267] J. H. Holland. Adaptation in natural and artificial systems. *University of Michigan Press*, 1975. (cited p. 203.)

[268] C. Darwin. On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. *New York: D. Appleton*. (cited p. 204.)

[269] I. N. Serdyuk, N. R. Zaccai, and G. Zaccaï. Methods in molecular biophysics: structure, dynamics, function. *Cambridge University Press*, 2007. (cited p. 206.)

[270] X. Chen, L. Sagle, and P. Cremer. Urea orientation at protein surfaces. *J Am Chem Soc*, 129(49):15104–15105, 2007. (cited p. 215.)

[271] F. Gabel, M. Jensen, G. Zaccaï, and M. Blackledge. Quantitative modelfree analysis of urea binding to unfolded ubiquitin using a combination of small angle x-ray and neutron scattering. *J Am Chem Soc*, 131(25):8769–8771, 2009. (cited p. 215.)

[272] H. Zhang, S. Neal, and D. S. Wishart. Refdb: a database of uniformly referenced protein chemical shifts. *J Biomol NMR*, 25(3):173–195, 2003. (cited p. 216.)

[273] E. Eyal, R. Najmanovich, B. Mcconkey, M. Edelman, and V. Sobolev. Importance of solvent accessibility and contact surfaces in modeling side-chain conformations in proteins. *J comput chem*, 25(5):712–724, 2004. (cited p. 217.)

[274] Y. Shen and A. Bax. Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. *J Biomol NMR*, 38 (4):289–302, 2007. (cited p. 217 and 218.)

[275] N. Maiti, M. Apetri, M. Zagorski, P. Carey, and V. Anderson. Raman spectroscopic characterization of secondary structure in natively unfolded proteins: alpha-synuclein. *J Am Chem Soc*, 126(8):2399–2408, 2004. (cited p. 220.)

[276] Z. Shi, K. Chen, Z. Liu, and N. Kallenbach. Conformation of the backbone in unfolded proteins. *Chem Rev*, 106(5):1877–1897, 2006. (cited p. 220.)

[277] K. Houben, L. Blanchard, M. Blackledge, and D. Marion. Intrinsic dynamics of the partly unstructured px domain from the sendai virus rna polymerase cofactor p. *Biophys J*, 93(8):2830–2844, 2007. (cited p. 221.)

[278] D. Sezer, J. Freed, and B. Roux. Parametrization, molecular dynamics simulation, and calculation of electron spin resonance spectra of a nitroxide spin label on a polyalanine alpha-helix. *J. Phys. Chem. B*, 112(18):5755–5767, 2008. (cited p. 224 and 226.)

[279] I. Solomon. Relaxation processes in a system of two spins. *Phys Rev*, 99(2): 559–566, 1955. (cited p. 225.)

[280] N. Bloembergen and L. Morgan. Proton relaxation times in paramagnetic solutions effects of electron spin relaxation. *J. Chem. Phys.*, 34(3):842, 1971. (cited p. 225.)

[281] R. Brüschweiler, B. Roux, M. Blackledge, C. Griesinger, M. Karplus, and R. Ernst. Influence of rapid intramolecular motion on nmr cross-relaxation rates. a molecular dynamics study of antamanide in solution. *J Am Chem Soc*, 114(7):2289–2302, 1992. (cited p. 225 and 226.)

[282] G. M. Clore and J. Iwahara. Theory, practice, and applications of paramagnetic relaxation enhancement for the characterization of transient low-population states of biological macromolecules and their complexes. *Chem Rev*, 109(9): 4108–39, 2009. (cited p. 225 and 227.)

[283] M. Cho, G. Nodet, H. Kim, M. Jensen, P. Bernado, C. Fernandez, S. Becker, M. Blackledge, and M. Zweckstetter. Structural characterization of alpha-synuclein in an aggregation prone state. *Protein Sci*, 18(9):1840–1846, 2009. (cited p. 229.)

[284] D. Ganguly and J. Chen. Structural interpretation of paramagnetic relaxation enhancement-derived distances for disordered protein states. *J. Mol. Biol.*, 390(3):467–477, 2009. (cited p. 233.)

[285] T. S. Kuhn. The structure of scientific revolutions. *University of Chicago Press*, 1962. (cited p. 253.)

[286] A. Quarteroni, R. Sacco, and F. Saleri. Méthodes numériques: algorithmes, analyse et applications. *Springer*, 2007. (cited p. 258.)

[287] A. Dermanis, A. Grün, and F. Sansò. Geomatic methods for the analysis of data in the earth sciences. *Springer*, 2000. (cited p. 258.)

[288] M. Giaquinta and G. Modica. Mathematical analysis: linear and metric structures and continuity. *Birkhauser*, page 465, 2007. (cited p. 258.)

[289] J. Losonczi, M. Andrec, M. Fischer, and J. Prestegard. Order matrix analysis of residual dipolar couplings using singular value decomposition. *J Magn Res*, 138(2):334–342, 1999. (cited p. 259.)

[290] J. Hurley, S. Lee, and G. Prag. Ubiquitin-binding domains. *Biochem J*, 399:361, 2006. (cited p. 261.)

[291] A. Hershko, A. Ciechanover, and A. Varshavsky. The ubiquitin system. *Nat med*, 6(10):1073–1081, 2000. (cited p. 261.)

[292] Y. Ye and M. Rape. Building ubiquitin chains: E2 enzymes at work. *Nature Reviews Molecular Cell Biology*, 10:755–764, 2009. (cited p. 261.)

[293] B. Mayer. Sh3 domains: complexity in moderation. *J Cell Sci*, 114(7):1253, 2001. (cited p. 261.)

[294] S. M. Larson and A. R. Davidson. The identification of conserved interactions within the sh3 domain by alignment of sequences and structures. *Protein Sci*, 9(11):2170–80, 2000. (cited p. 261.)

[295] S. Stamenova, M. French, Y. He, S. Francis, Z. Kramer, and L. Hicke. Ubiquitin binds to and regulates a subset of sh3 domains. *Mol Cell*, 25(2):273–284, 2007. (cited p. 261.)

[296] M. Skiadopoulos, S. Surman, J. Riggs, W. Elkins, M. S. Claire, M. Nishio, D. Garcin, D. Kolakofsky, P. Collins, and B. Murphy. Sendai virus, a murine parainfluenza virus type 1, replicates to a level similar to human piv1 in the upper and lower respiratory tract of african green monkeys and chimpanzees. *Virology*, 297(1):153–160, 2002. (cited p. 261.)

[297] D. Kolakofsky, P. L. Mercier, F. Iseni, and D. Garcin. Viral rna polymerase scanning and the gymnastics of sendai virus rna synthesis. *Virology*, 318(2): 463–473, 2004. (cited p. 261.)

[298] M. Goedert. Alpha-synuclein and neurodegenerative diseases. *Nature Reviews Neuroscience*, 2(7):492–501, 2001. (cited p. 261.)

[299] M. Spillantini, M. Schmidt, M. Virginia, J. Trojanowski, R. Jakes, and M. Goedert. alpha-synuclein in lewy bodies. *Nature*, 388(6645):839–840, 1997. (cited p. 261.)

[300] K. Uéda, H. Fukushima, E. Masliah, Y. Xia, A. Iwai, M. Yoshimoto, D. Otero, J. Kondo, Y. Ihara, and T. Saitoh. Molecular cloning of cdna encoding an unrecognized component of amyloid in alzheimer disease. *Proc Natl Acad Sci USA*, 90(23):11282, 1993. (cited p. 261.)

[301] P. Lansbury. Genetics of parkinson's disease and biochemical studies of implicated gene products: Commentary. *Curr op cell biol*, 14(5):653–660, 2002. (cited p. 261.)

RÉSUMÉ Les macromolécules biologiques sont, par essence, des systèmes dynamiques. Si l'importance de cette flexibilité est maintenant clairement établie, la caractérisation précise du désordre conformationnel de ces systèmes reste encore une question ouverte. La résonance magnétique nucléaire constitue un outil unique pour sonder ces mouvements au niveau atomique que ce soit par les études de relaxation de spin ou par l'analyse des couplages dipolaires résiduels. Ces derniers permettent d'étudier l'ensemble des mouvements ayant lieu à des échelles de temps plus rapide que la milliseconde, englobant ainsi les temps caractéristiques de nombreux mouvements physiologiquement importants. L'information contenue dans ces couplages résiduels est ici interprétée principalement grâce à des approches analytiques pour quantifier la dynamique présente dans des protéines repliées, déterminer l'orientation de ces mouvements et obtenir de l'information structurale au sein de ce désordre conformationnel. Ces approches analytiques sont complémentées par des méthodes numériques, permettant ainsi soit d'observer les phénomènes sous un autre angle, soit d'examiner d'autres systèmes tels que les protéines intrinsèquement désordonnées. L'ensemble de ces études laisse transparaître une importante complémentarité entre ordre structural et désordre conformationnel.

MOTS-CLÉS Résonance Magnétique Nucléaire, Couplages Dipolaires Résiduels, Désordre Conformationnel, Protéines, Structure, Dynamique.

ABSTRACT Biological macromolecules are, by essence, dynamical systems. While the importance of this flexibility is nowadays well established, the accurate characterization of the conformational disorder of these systems remains an important challenge. Nuclear magnetic resonance spectroscopy is a unique tool to probe these motions at atomic level, through the analysis of spin relaxation or residual dipolar couplings. The latter allows all motions occurring at timescales faster than the millisecond to be investigated, including physiologically important timescales. The information presents in those couplings is interpreted here using mainly analytical approaches in order to quantify the amounts of dynamics present in folded protein, to determine the direction of those motions and to obtain structural information within this conformational disorder. These analytical approaches are complemented by numerical methods, that allowed the observation of phenomena from a different point of view or the investigation of other systems such as intrinsically disordered proteins. All of these studies demonstrate an important complementarity between structural order and conformational disorder.

KEY WORDS Nuclear Magnetic Resonance, Residual Dipolar Coupling, Conformational Disorder, Protein, Dynamics, Structure.

LABORATOIRE DE THÈSE Institut de Biologie Structurale Jean-Pierre Ebel, UMR 5075, CEA-CNRS-UJF. Equipe Flexibilité et Dynamique des Protéines. 41, rue Jules Horowitz, 38027 Grenoble Cedex.